

# FACTOR MODELING FOR HIGH-DIMENSIONAL TIME SERIES: INFERENCE FOR THE NUMBER OF FACTORS<sup>1</sup>

BY CLIFFORD LAM AND QIWEI YAO

*London School of Economics and Political Science, and London School of  
 Economics and Political Science and Guanghua School of Management,  
 Peking University*

This paper deals with the factor modeling for high-dimensional time series based on a dimension-reduction viewpoint. Under stationary settings, the inference is simple in the sense that both the number of factors and the factor loadings are estimated in terms of an eigenanalysis for a nonnegative definite matrix, and is therefore applicable when the dimension of time series is on the order of a few thousands. Asymptotic properties of the proposed method are investigated under two settings: (i) the sample size goes to infinity while the dimension of time series is fixed; and (ii) both the sample size and the dimension of time series go to infinity together. In particular, our estimators for zero-eigenvalues enjoy faster convergence (or slower divergence) rates, hence making the estimation for the number of factors easier. In particular, when the sample size and the dimension of time series go to infinity together, the estimators for the eigenvalues are no longer consistent. However, our estimator for the number of the factors, which is based on the ratios of the estimated eigenvalues, still works fine. Furthermore, this estimation shows the so-called “blessing of dimensionality” property in the sense that the performance of the estimation may improve when the dimension of time series increases. A two-step procedure is investigated when the factors are of different degrees of strength. Numerical illustration with both simulated and real data is also reported.

**1. Introduction.** The analysis of multivariate time series data is of increased interest and importance in the modern information age. Although the methods and the associate theory for univariate time series analysis are

---

Received March 2011; revised January 2012.

<sup>1</sup>Supported in part by the Engineering and Physical Sciences Research Council of the United Kingdom.

*AMS 2000 subject classifications.* Primary 62M10, 62H30; secondary 60G99.

*Key words and phrases.* Autocovariance matrices, blessing of dimensionality, eigenanalysis, fast convergence rates, multivariate time series, ratio-based estimator, strength of factors, white noise.

This is an electronic reprint of the original article published by the  
 Institute of Mathematical Statistics in *The Annals of Statistics*,  
 2012, Vol. 40, No. 2, 694–726. This reprint differs from the original in pagination  
 and typographic detail.

well developed and understood, the picture for the multivariate cases is less complete. In spite of the fact that the conventional univariate time series models (such as ARMA) and the associated time-domain and frequency-domain methods have been formally extended to multivariate cases, their usefulness is often limited. One may face serious issues such as the lack of model identification or flat likelihood functions. In fact vector ARMA models are seldom used directly in practice. Dimension-reduction via, for example, reduced-rank structure, structural indices, scalar component models and canonical correlation analysis is more pertinent in modeling multivariate time series data. See [10, 14, 20, 22].

In this paper we deal with the factor modeling for multivariate time series from a dimension-reduction viewpoint. Differently from the factor analysis for independent observations, we search for the factors which drive the serial dependence of the original time series. Early attempts in this direction include [1, 5, 16, 18, 21, 23, 25]. More recent efforts focus on the inference when the dimension of time series is as large as or even greater than the sample size; see, for example, [13] and the references within. High-dimensional time series data are often encountered nowadays in many fields including finance, economics, environmental and medical studies. For example, understanding the dynamics of the returns of large numbers of assets is the key for asset pricing, portfolio allocation, and risk management. Panel time series are commonplace in studying economic and business phenomena. Environmental time series are often of a high dimension due to a large number of indices monitored across many different locations.

Our approach is from a dimension-reduction point of view. The model adopted can be traced back at least to that of [18]. We decompose a high-dimensional time series into two parts: a dynamic part driven by, hopefully, a lower-dimensional factor time series, and a static part which is a vector white noise. Since the white noise exhibits no serial correlations, the decomposition is unique in the sense that both the number of factors (i.e., the dimension of the factor process) and the factor loading space in our model are identifiable. Such a conceptually simple decomposition also makes the statistical inference easy. Although the setting allows the factor process to be nonstationary (see [16]; also Section 2.1 below), we focus on stationary models only in this paper: under the stationary condition, the estimation for both the number of factors and the factor loadings is carried out in an eigenanalysis for a nonnegative definite matrix, and is therefore applicable when the dimension of time series is on the order of a few thousands. Furthermore, the asymptotic properties of the proposed method are investigated under two settings: (i) the sample size goes to infinity while the dimension of time series is fixed; and (ii) both the sample size and the dimension of time series go to infinity together. In particular, our estimators for zero-eigenvalues enjoy the faster convergence (or slower divergence) rates, from

which the proposed ratio-based estimator for the number of factors benefits. In fact when all the factors are strong, the performance of our estimation for the number of factors improves when the dimension of time series increases. This phenomenon is coined as “blessing of dimensionality.”

The new contributions of this paper include: (i) the ratio-based estimator for the number of factors and the associated asymptotic theory which underpins the “blessing of dimensionality” phenomenon observed in numerical experiments, and (ii) a two-step estimation procedure when the factors are of different degrees of strength. We focus on the results related to the estimation for the number of factors in this paper. The results on the estimation of the factor loading space under the assumption that the number of factors is known are reported in [13].

There exists a large body of literature in econometrics and finance on factor models for high-dimensional time series. However, most of them are based on a different viewpoint, as those models attempt to identify the *common factors* that affect the dynamics of most original component series. In analyzing economic and financial phenomena, it is often appealing to separate these common factors from the so-called idiosyncratic components: each idiosyncratic component may at most affect the dynamics of a few original time series. An idiosyncratic series may exhibit serial correlations and, therefore, may be a time series itself. This poses technical difficulties in both model identification and inference. In fact the rigorous definition of the common factors and the idiosyncratic components can only be established asymptotically when the dimension of time series tends to infinity; see [6, 8]. Hence those factor models are only asymptotically identifiable. According to the definition adopted in this paper, both “the common factors” and those serially correlated idiosyncratic components will be identified as factors. This is not ideal for the applications with the purpose to identify those common factors. However, this makes the tasks of model identification and inference much simpler.

The rest of the paper is organized as follows. The model and the estimation methods are introduced in Section 2. The sampling properties of the estimation methods are investigated in Section 3. Simulation results are inserted whenever appropriate to illustrate the various asymptotic properties of the methods. Section 4 deals with the cases when different factors are of different strength, for which a two-step estimation procedure is preferred. The analysis of two real data sets is reported in Section 5. All mathematical proofs are relegated to the [Appendix](#).

## 2. Models and estimation.

2.1. *Models.* If we are interested in the linear dynamic structure of  $\mathbf{y}_t$  only, conceptually we may think that  $\mathbf{y}_t$  consists of two parts: a static part (i.e., a white noise), and a dynamic component driven by, hopefully, a low-

dimensional process. This leads to the decomposition:

$$(2.1) \quad \mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

where  $\mathbf{x}_t$  is an  $r \times 1$  latent process with (unknown)  $r \leq p$ ,  $\mathbf{A}$  is a  $p \times r$  unknown constant matrix, and  $\boldsymbol{\varepsilon}_t \sim \text{WN}(\boldsymbol{\mu}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)$  is a vector white-noise process. When  $r$  is much smaller than  $p$ , we achieve an effective dimension-reduction, as then the serial dependence of  $\mathbf{y}_t$  is driven by that of a much lower-dimensional process  $\mathbf{x}_t$ . We call  $\mathbf{x}_t$  a factor process. The setting (2.1) may be traced back at least to [18]; see also its further development in dealing with cointegrated factors in [19].

Since none of the elements on the RHS of (2.1) are observable, we have to characterize them further to make them identifiable. First we assume that no linear combinations of  $\mathbf{x}_t$  are white noise, as any such components can be absorbed into  $\boldsymbol{\varepsilon}_t$  [see condition (C1) below]. We also assume that the rank of  $\mathbf{A}$  is  $r$ . Otherwise (2.1) may be expressed equivalently in terms of a lower-dimensional factor. Furthermore, since (2.1) is unchanged if we replace  $(\mathbf{A}, \mathbf{x}_t)$  by  $(\mathbf{A}\mathbf{H}, \mathbf{H}^{-1}\mathbf{x}_t)$  for any invertible  $r \times r$  matrix  $\mathbf{H}$ , we may assume that the columns of  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_r)$  are orthonormal, that is,  $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$ , where  $\mathbf{I}_r$  denotes the  $r \times r$  identity matrix. Note that even with this constraint,  $\mathbf{A}$  and  $\mathbf{x}_t$  are not uniquely determined in (2.1), as the aforementioned replacement is still applicable for any orthogonal  $\mathbf{H}$ . However, the factor loading space, that is, the  $r$ -dimensional linear space spanned by the columns of  $\mathbf{A}$ , denoted by  $\mathcal{M}(\mathbf{A})$ , is uniquely defined.

We summarize into condition (C1) all the assumptions introduced so far:

(C1) In model (2.1),  $\boldsymbol{\varepsilon}_t \sim \text{WN}(\boldsymbol{\mu}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)$ . If  $\mathbf{c}'\mathbf{X}_t$  is white noise for a constant  $\mathbf{c} \in \mathbb{R}^p$ , then  $\mathbf{c}'\text{Cov}(\mathbf{X}_{t+k}, \boldsymbol{\varepsilon}_t) = \mathbf{0}$  for any nonzero integers  $k$ . Furthermore  $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$ .

The key for the inference for model (2.1) is to determine the number of factors  $r$  and to estimate the  $p \times r$  factor loading matrix  $\mathbf{A}$ , or more precisely the factor loading space  $\mathcal{M}(\mathbf{A})$ . Once we have obtained an estimator, say,  $\hat{\mathbf{A}}$ , a natural estimator for the factor process is

$$(2.2) \quad \hat{\mathbf{x}}_t = \hat{\mathbf{A}}'\mathbf{y}_t,$$

and the resulting residuals are

$$(2.3) \quad \hat{\boldsymbol{\varepsilon}}_t = (\mathbf{I}_d - \hat{\mathbf{A}}\hat{\mathbf{A}}')\mathbf{y}_t.$$

The dynamic modeling for  $\mathbf{y}_t$  is achieved via such a modeling for  $\hat{\mathbf{x}}_t$  and the relationship  $\hat{\mathbf{y}}_t = \hat{\mathbf{A}}\hat{\mathbf{x}}_t$ . A parsimonious fitting for  $\hat{\mathbf{x}}_t$  may be obtained by rotating  $\hat{\mathbf{x}}_t$  appropriately [27]. Such a rotation is equivalent to replacing  $\hat{\mathbf{A}}$  by  $\hat{\mathbf{A}}\mathbf{H}$  for an appropriate  $r \times r$  orthogonal matrix  $\mathbf{H}$ . Note that  $\mathcal{M}(\hat{\mathbf{A}}) = \mathcal{M}(\hat{\mathbf{A}}\mathbf{H})$ , and the residuals (2.3) are unchanged with such a replacement.

2.2. *Estimation for  $\mathbf{A}$  and  $r$ .* An innovation expansion algorithm is proposed in [16] for estimating  $\mathbf{A}$  based on solving a sequence of nonlinear optimization problems with at most  $p$  variables. Although the algorithm is feasible for small or moderate  $p$  only, it can handle the situations when the factor process  $\mathbf{x}_t$  is nonstationary. We outline the key idea below, as our computationally more efficient estimation method for stationary cases is based on the same principle.

Our goal is to estimate  $\mathcal{M}(\mathbf{A})$ , or, equivalently, its orthogonal complement  $\mathcal{M}(\mathbf{B})$ , where  $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_{p-r})$  is a  $p \times (p-r)$  matrix for which  $(\mathbf{A}, \mathbf{B})$  forms a  $p \times p$  orthogonal matrix, that is,  $\mathbf{B}'\mathbf{A} = 0$  and  $\mathbf{B}'\mathbf{B} = \mathbf{I}_{p-r}$  [see also (C1)]. It follows from (2.1) that

$$(2.4) \quad \mathbf{B}'\mathbf{y}_t = \mathbf{B}'\boldsymbol{\varepsilon}_t,$$

implying that for any  $1 \leq j \leq p-r$ ,  $\{\mathbf{b}_j'\mathbf{y}_t, t = 0, \pm 1, \dots\}$  is a white-noise process. Hence, we may search for mutually orthogonal directions  $\mathbf{b}_1, \mathbf{b}_2, \dots$  one by one such that the projection of  $\mathbf{y}_t$  on each of those directions is a white noise. We stop the search when such a direction is no longer available, and take  $p-k$  as the estimated value of  $r$ , where  $k$  is the number of directions obtained in the search. This is essentially how [16] accomplish the estimation. It is irrelevant in the above derivation if  $\mathbf{x}_t$  is stationary or not.

However, a much simpler method is available when  $\mathbf{x}_t$ , therefore also  $\mathbf{y}_t$ , is stationary:

(C2)  $\mathbf{x}_t$  is weakly stationary, and  $\text{Cov}(\mathbf{x}_t, \boldsymbol{\varepsilon}_{t+k}) = 0$  for any  $k \geq 0$ .

In most factor modeling literature,  $\mathbf{x}_t$  and  $\boldsymbol{\varepsilon}_s$  are assumed to be uncorrelated for any  $t$  and  $s$ . Condition (C2) requires only that the future white-noise components are uncorrelated with the factors up to the present. This enlarges the model capacity substantially. Put

$$\begin{aligned} \boldsymbol{\Sigma}_y(k) &= \text{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t), & \boldsymbol{\Sigma}_x(k) &= \text{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t), \\ \boldsymbol{\Sigma}_{x\varepsilon}(k) &= \text{Cov}(\mathbf{x}_{t+k}, \boldsymbol{\varepsilon}_t). \end{aligned}$$

It follows from (2.1) and (C2) that

$$(2.5) \quad \boldsymbol{\Sigma}_y(k) = \mathbf{A}\boldsymbol{\Sigma}_x(k)\mathbf{A}' + \mathbf{A}\boldsymbol{\Sigma}_{x\varepsilon}(k), \quad k \geq 1.$$

For a prescribed integer  $k_0 \geq 1$ , define

$$(2.6) \quad \mathbf{M} = \sum_{k=1}^{k_0} \boldsymbol{\Sigma}_y(k) \boldsymbol{\Sigma}_y(k)'. \quad \mathbf{M} = \sum_{k=1}^{k_0} \boldsymbol{\Sigma}_y(k) \boldsymbol{\Sigma}_y(k)'.$$

Then  $\mathbf{M}$  is a  $p \times p$  nonnegative matrix. It follows from (2.5) that  $\mathbf{M}\mathbf{B} = 0$ , that is, the columns of  $\mathbf{B}$  are the eigenvectors of  $\mathbf{M}$  corresponding to zero-eigenvalues. Hence conditions (C1) and (C2) imply:

*The factor loading space  $\mathcal{M}(\mathbf{A})$  is spanned by the eigenvectors of  $\mathbf{M}$  corresponding to its nonzero eigenvalues, and the number of the nonzero eigenvalues is  $r$ .*

We take the sum in the definition of  $\mathbf{M}$  to accumulate the information from different time lags. This is useful especially when the sample size  $n$  is small. We use the nonnegative definite matrix  $\boldsymbol{\Sigma}_y(k)\boldsymbol{\Sigma}_y(k)'$  [instead of  $\boldsymbol{\Sigma}_y(k)$ ] to avoid the cancellation of the information from different lags. This is guaranteed by the fact that for any matrix  $\mathbf{C}$ ,  $\mathbf{M}\mathbf{C} = 0$  if and only if  $\boldsymbol{\Sigma}_y(k)'\mathbf{C} = 0$  for all  $1 \leq k \leq k_0$ . We tend to use small  $k_0$ , as the autocorrelation is often at its strongest at the small time lags. On the other hand, adding more terms will not alter the value of  $r$ , although the estimation for  $\boldsymbol{\Sigma}_y(k)$  with large  $k$  is less accurate. The simulation results reported in [13] also confirm that the estimation for  $\mathbf{A}$  and  $r$ , defined below, is not sensitive to the choice of  $k_0$ .

To estimate  $\mathcal{M}(\mathbf{A})$ , we only need to perform an eigenanalysis on

$$(2.7) \quad \widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\boldsymbol{\Sigma}}_y(k) \widehat{\boldsymbol{\Sigma}}_y(k)',$$

where  $\widehat{\boldsymbol{\Sigma}}_y(k)$  denotes the sample covariance matrix of  $\mathbf{y}_t$  at lag  $k$ . Then the estimator  $\widehat{r}$  for the number of factors is defined in (2.8) below. The columns of the estimated factor loading matrix  $\widehat{\mathbf{A}}$  are the  $\widehat{r}$  orthonormal eigenvectors of  $\widehat{\mathbf{M}}$  corresponding to its  $\widehat{r}$  largest eigenvalues. Note that the estimator  $\widehat{\mathbf{A}}$  is essentially the same as that defined in Section 2.4 of [13], although a canonical form of the model is used there in order to define the factor loading matrix uniquely.

Due to the random fluctuation in a finite sample, the estimates for the zero-eigenvalues of  $\mathbf{M}$  are unlikely to be 0 exactly. A common practice is to plot all the estimated eigenvalues in a descending order, and look for a cut-off value  $\widehat{r}$  such that the  $(\widehat{r} + 1)$ th largest eigenvalue is substantially smaller than the  $\widehat{r}$  largest eigenvalues. This is effectively an eyeball-test. The ratio-based estimator defined below may be viewed as an enhanced eyeball-test, based on the same idea as [28]. In fact this ratio-based estimator benefits from the faster convergence rates of the estimators for the zero-eigenvalues; see Proposition 1 in Section 3.1 below, and also Theorems 1 and 2 in Section 3.2 below. The other available methods for determining  $r$  include the information criteria approaches of [2, 3] and [9], and the bootstrap approach of [4], though the settings considered in those papers are different.

*A ratio-based estimator for  $r$ .* We define an estimator for the number of factors  $r$  as follows:

$$(2.8) \quad \widehat{r} = \arg \min_{1 \leq i \leq R} \widehat{\lambda}_{i+1} / \widehat{\lambda}_i,$$

where  $\widehat{\lambda}_1 \geq \dots \geq \widehat{\lambda}_p$  are the eigenvalues of  $\widehat{\mathbf{M}}$ , and  $r < R < p$  is a constant.

In practice we may use, for example,  $R = p/2$ . We cannot extend the search up to  $p$ , as the minimum eigenvalue of  $\widehat{\mathbf{M}}$  is likely to be practically 0,

especially when  $n$  is small and  $p$  is large. It is worthy noting that when  $p$  and  $n$  are on the same order, the estimators for eigenvalues are no longer consistent. However, the ratio-based estimator (2.8) still works well. See Theorem 2(iii) below.

The above estimation methods for  $\mathbf{A}$  and  $r$  can be extended to those non-stationary time series for which a generalized lag- $k$  autocovariance matrix is well defined (see, e.g., [19]). In fact, the methods are still applicable when the weak limit of the generalized lag- $k$  autocovariance matrix

$$\widehat{\mathbf{S}}_y(k) = n^{-\alpha} \sum_{t=1}^{n-1} (\mathbf{y}_{t+k} - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})'$$

exists for  $1 \leq k \leq k_0$ , where  $\alpha > 1$  is a constant. Further developments on those lines will be reported elsewhere. For the factor modeling for high-dimensional volatility processes based on a similar idea, we refer to [15, 26].

**3. Estimation properties.** Conventional asymptotic properties are established under the setting that the sample size  $n$  tends to  $\infty$  and everything else remains fixed. Modern time series analysis encounters the situation when the number of time series  $p$  is as large as, or even larger than, the sample size  $n$ . Then the asymptotic properties established under the setting when both  $n$  and  $p$  tend to  $\infty$  are more relevant. We deal with these two settings in Section 3.1 and Sections 3.2–3.4 separately.

**3.1. Asymptotics when  $n \rightarrow \infty$  and  $p$  fixed.** We first consider the asymptotic properties under the assumption that  $n \rightarrow \infty$  and  $p$  is fixed. These properties reflect the behavior of our estimation method in the cases when  $n$  is large and  $p$  is small. We introduce some regularity conditions first. Let  $\lambda_1, \dots, \lambda_p$  be the eigenvalues of the matrix  $\mathbf{M}$ :

(C3)  $\mathbf{y}_t$  is strictly stationary and  $\psi$ -mixing with the mixing coefficients  $\psi(\cdot)$  satisfying the condition that  $\sum_{t \geq 1} t\psi(t)^{1/2} < \infty$ . Furthermore,  $E\{|\mathbf{y}_t|^4\} < \infty$  element-wisely.

(C4)  $\lambda_1 > \dots > \lambda_r > 0 = \lambda_{r+1} = \dots = \lambda_p$ .

Section 2.6 of [7] gives a compact survey on the mixing properties of time series. The use of the  $\psi$ -mixing condition in (C3) is for technical convenience. Note that  $\mathbf{M}$  is a nonnegative definite matrix. All its eigenvalues are nonnegative. Condition (C4) assumes that its  $r$  nonzero eigenvalues are distinct from each other. While this condition is not essential, it substantially simplifies the presentation of the convergence properties in Proposition 1 below. Let  $\boldsymbol{\gamma}_j$  be a unit eigenvector of  $\mathbf{M}$  corresponding to the eigenvalue  $\lambda_j$ . We denote by  $(\widehat{\lambda}_1, \widehat{\boldsymbol{\gamma}}_1), \dots, (\widehat{\lambda}_p, \widehat{\boldsymbol{\gamma}}_p)$  the  $p$  pairs of eigenvalue and eigenvector of matrix  $\widehat{\mathbf{M}}$ : the eigenvalues  $\widehat{\lambda}_j$  are arranged in descending order, and the



eigenvectors  $\hat{\gamma}_j$  are orthonormal. Furthermore, it may go without explicit statement that  $\hat{\gamma}_j$  may be replaced by  $-\hat{\gamma}_j$  in order to match the direction of  $\gamma_j$  for  $1 \leq j \leq r$ .

**PROPOSITION 1.** *Let conditions (C1)–(C4) hold. Then as  $n \rightarrow \infty$  (but  $p$  fixed), it holds that:*

- (i)  $|\hat{\lambda}_j - \lambda_j| = O_P(n^{-1/2})$  and  $\|\hat{\gamma}_j - \gamma_j\| = O_P(n^{-1/2})$  for  $j = 1, \dots, r$ , and
- (ii)  $\hat{\lambda}_j = O_P(n^{-1})$  for  $j = r+1, \dots, p$ .

The proof of the above proposition is in principle the same as that of Theorem 1 in [4], and is therefore omitted.

**3.2. Asymptotics when  $n \rightarrow \infty, p \rightarrow \infty$  and  $r$  fixed.** To highlight the radically different behavior when  $p$  diverges together with  $n$ , we first conduct some simulations: we set in model (2.1)  $r = 1$ ,  $\mathbf{A}' = (1, \dots, 1)$ ,  $\varepsilon_t$  are independent  $N(0, \mathbf{I}_p)$ , and  $\mathbf{x}_t = x_t$  is an AR(1) process defined by  $x_{t+1} = 0.7x_t + e_t$ . We set the sample size  $n = 50, 100, 200, 400, 800, 1600$  and 3200, and the dimension fixed at half the sample size, that is,  $p = n/2$ . Let  $\mathbf{M}$  be defined as in (2.6) with  $k_0 = 1$ . For each setting, we draw 200 samples. The boxplots of the errors  $\hat{\lambda}_i - \lambda_i$ ,  $i = 1, \dots, 6$ , are depicted in Figure 1. Note that  $\lambda_i = 0$  for  $i \geq 2$ , since  $r = 1$ . The figure shows that those estimation errors do not converge to 0. In fact those errors seem to increase when  $n$  (and also  $p = n/2$ ) increases. Therefore the classical asymptotic theory (i.e.,  $n \rightarrow \infty$  and  $p$  fixed) such as Proposition 1 above is irrelevant when  $p$  increases together with  $n$ . In spite of the lack of consistency in estimating the eigenvalues, the ratio-based estimator for the number of factors  $r$  ( $=1$ ) defined in (2.8) works perfectly fine for this example, as shown in Figure 2. In fact it is always the case that  $\hat{r} \equiv 1$  in all our experiments even when the sample size is as small as  $n = 50$ ; see Figure 2.

To develop the relevant asymptotic theory, we introduce some notation first. For any matrix  $\mathbf{G}$ , let  $\|\mathbf{G}\|$  be the square root of the maximum eigenvalue of  $\mathbf{G}\mathbf{G}'$ , and  $\|\mathbf{G}\|_{\min}$  be the square root of the smallest nonzero eigenvalue of  $\mathbf{G}\mathbf{G}'$ . We write  $a \asymp b$  if  $a = O(b)$  and  $b = O(a)$ . Recall  $\Sigma_x(k) = \text{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t)$  and  $\Sigma_{x\varepsilon}(k) = \text{Cov}(\mathbf{x}_{t+k}, \varepsilon_t)$ . Some regularity conditions are now in order:

- (C5) For a constant  $\delta \in [0, 1]$ , it holds that  $\|\Sigma_x(k)\| \asymp p^{1-\delta} \asymp \|\Sigma_x(k)\|_{\min}$ .
- (C6) For  $k = 0, 1, \dots, k_0$ ,  $\|\Sigma_{x\varepsilon}(k)\| = o(p^{1-\delta})$ .

**REMARK 1.** (i) Condition (C5) looks unnatural. It is derived from more natural conditions (3.1) and (3.2) below coupled with the standardization



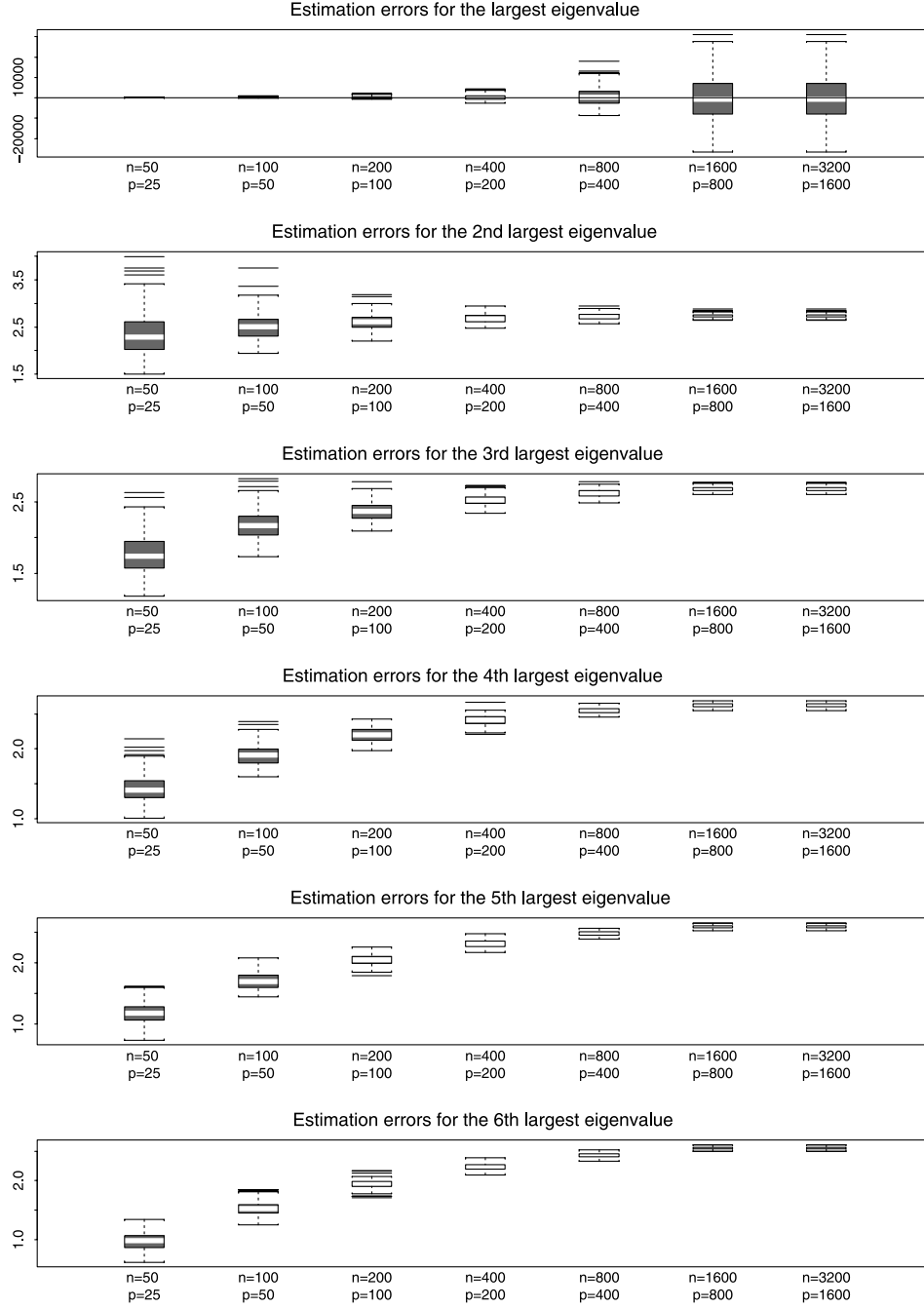


FIG. 1. *Boxplots for the errors in estimating the first six eigenvalues of  $\mathbf{M}$  with  $r = 1$  and all the factor loading coefficients being 1.*

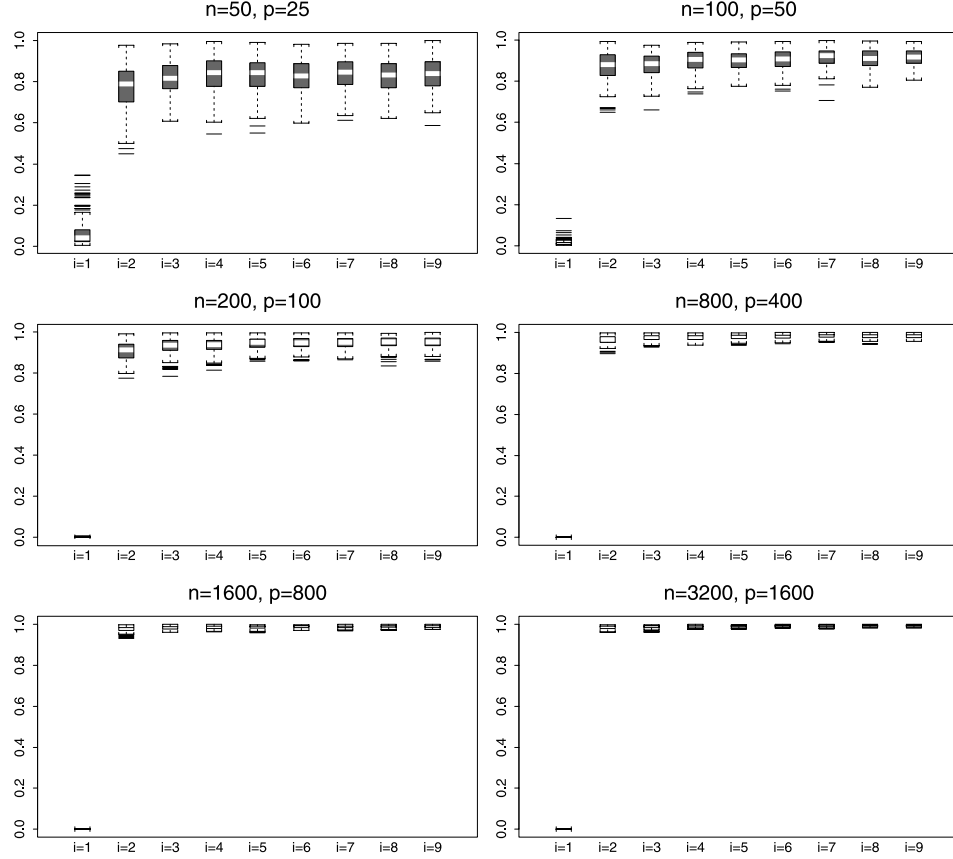


FIG. 2. Boxplots for the ratios  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$ , with  $r = 1$  and all the factor loading coefficients being 1.

$\mathbf{A}'\mathbf{A} = \mathbf{I}_r$ . Since  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_r)$  is  $p \times r$  and  $p \rightarrow \infty$  now, it is natural to let the norm of each column of  $\mathbf{A}$ , before standardizing to  $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$ , tend to  $\infty$  as well. To this end, we assume that

$$(3.1) \quad \|\mathbf{a}_j\|^2 \asymp p^{1-\delta_j}, \quad j = 1, \dots, r,$$

where  $\delta_j \in [0, 1]$  are constants. We take  $\delta_j$  as a measure of the strength of the factor  $x_{tj}$ . We call  $x_{tj}$  a strong factor when  $\delta_j = 0$ , and a weak factor when  $\delta_j > 0$ . Since  $r$  is fixed, it is also reasonable to assume that for  $k = 0, 1, \dots, k_0$ ,

$$(3.2) \quad |\Sigma_x(k)| \neq 0.$$

Then condition (C5) is entailed by the standardization  $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$  under conditions (3.2) and (3.1) with  $\delta_j = \delta$  for all  $j$ .

(ii) The condition assumed on  $\Sigma_{x\epsilon}(k)$  in (C6) requires that the correlation between  $\mathbf{x}_{t+k}$  ( $k \geq 0$ ) and  $\epsilon_t$  is not too strong. In fact under a natural

condition that  $\Sigma_{x\epsilon}(k) = O(1)$  element-wisely, it is implied by (3.1) and the standardization  $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$  [hence now  $x_{j,t} = O_P(p^{(1-\delta)/2})$  as a result of such standardization] that  $\|\Sigma_{x\epsilon}(k)\| = O(p^{1-\delta/2})$ .

Now we deal with the convergence rates of the estimated eigenvalues, and establish the results in the same spirit as Proposition 1. Of course the convergence (or divergence) rate for each estimator  $\hat{\lambda}_i$  is slower, as the number of estimated parameters goes to infinity now.

**THEOREM 1.** *Let conditions (C1)–(C6) hold and  $h_n = p^\delta n^{-1/2} \rightarrow 0$ . Then as  $n \rightarrow \infty$  and  $p \rightarrow \infty$ , it holds that:*

- (i)  $|\hat{\lambda}_i - \lambda_i| = O_P(p^{2-\delta} n^{-1/2})$  for  $i = 1, \dots, r$ , and
- (ii)  $\hat{\lambda}_j = O_P(p^2 n^{-1})$  for  $j = r+1, \dots, p$ .

**COROLLARY 1.** *Under the condition of Theorem 1, it holds that*

$$\hat{\lambda}_{j+1}/\hat{\lambda}_j \asymp 1 \quad \text{for } j = 1, \dots, r-1 \quad \text{and} \quad \hat{\lambda}_{r+1}/\hat{\lambda}_r = O_P(p^{2\delta}/n) \xrightarrow{P} 0.$$

The proofs of Theorem 1 and Corollary 1 are presented in the Appendix. Obviously when  $p$  is fixed, Theorem 1 formally reduces to Proposition 1. Some remarks are now in order.

**REMARK 2.** (i) Corollary 1 implies that the plot of ratios  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$ ,  $i = 1, 2, \dots$ , will drop sharply at  $i = r$ . This provides a partial theoretical underpinning for the estimator  $\hat{r}$  defined in (2.8). Especially when all factors are strong (i.e.,  $\delta = 0$ ),  $\hat{\lambda}_{r+1}/\hat{\lambda}_r = O_P(n^{-1})$ . This convergence rate is independent of  $p$ , suggesting that the estimation for  $r$  may not suffer as  $p$  increases. In fact when all the factors are strong, the estimation for  $r$  may improve as  $p$  increases. See Remark 3(iv) in Section 3.4 below.

(ii) Unfortunately, we are unable to derive an explicit asymptotic expression for the ratios  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$  with  $i > r$ , although we make the following conjecture:

$$(3.3) \quad \hat{\lambda}_{j+1}/\hat{\lambda}_j \xrightarrow{P} 1, \quad j = (k_0 + 1)r + 1, \dots, (k_0 + 1)r + K,$$

where  $k_0$  is the number of lags used in defining matrix  $\mathbf{M}$  in (2.6), and  $K \geq 1$  is any fixed integer. See also Figure 2. Further simulation results, not reported explicitly, also conform with (3.3). This conjecture arises from the following observation: for  $j > (k_0 + 1)r$ , the  $j$ th largest eigenvalue of  $\hat{\mathbf{M}}$  is predominately contributed by the term  $\sum_{k=1}^{k_0} \hat{\Sigma}_\varepsilon(k) \hat{\Sigma}_\varepsilon(k)'$  which has a cluster of largest eigenvalues on the order of  $p^2/n^2$ , where  $\hat{\Sigma}_\varepsilon(k)$  is the sample lag- $k$  autocovariance matrix for  $\varepsilon_t$ . See also Theorem 2(iii) in Section 3.4 below.

TABLE 1  
Relative frequency estimates for  $P(\hat{r}=r)$  in the simulation with 200 replications

	$n$	50	100	200	400	800	1600	3200
$\delta = 0$	$p = 0.2n$	0.165	0.680	0.940	0.995	1	1	1
	$p = 0.5n$	0.410	0.800	0.980	1	1	1	1
	$p = 0.8n$	0.560	0.815	0.990	1	1	1	1
	$p = 1.2n$	0.590	0.820	0.990	1	1	1	1
$\delta = 0.5$	$p = 0.2n$	0.075	0.155	0.270	0.570	0.980	1	1
	$p = 0.5n$	0.090	0.285	0.285	0.820	0.960	1	1
	$p = 0.8n$	0.060	0.180	0.490	0.745	0.970	1	1
	$p = 1.2n$	0.090	0.180	0.310	0.760	0.915	1	1

(iii) The errors in estimating eigenvalues are on the order of  $p^{2-\delta}n^{-1/2}$  or  $p^2n^{-1}$ , and both do not necessarily converge to 0. However, since

$$\frac{\hat{\lambda}_j}{|\hat{\lambda}_i - \lambda_i|} = O_P(p^\delta n^{-1/2}) = O_P(h_n) = o_P(1)$$

for any  $1 \leq i \leq r$  and  $r < j \leq p$ ,

the estimation errors for the zero-eigenvalues is asymptotically of an order of magnitude smaller than those for the nonzero-eigenvalues.

**3.3. Simulation.** To illustrate the asymptotic properties in Section 3.2 above, we report some simulation results. We set in model (2.1)  $r = 3$ ,  $n = 50, 100, 200, 400, 800, 1600$  and  $3200$ , and  $p = 0.2n, 0.5n, 0.8n$  and  $1.2n$ . All the  $p \times r$  elements of  $\mathbf{A}$  are generated independently from the uniform distribution on the interval  $[-1, 1]$  first, and we then divide each of them by  $p^{\delta/2}$  to make all three factors of the strength  $\delta$ ; see (3.1). We generate factor  $\mathbf{x}_t$  from a  $3 \times 1$  vector-AR(1) process with independent  $N(0, 1)$  innovations and the diagonal autoregressive coefficient matrix with 0.6,  $-0.5$  and  $0.3$  as the main diagonal elements. We let  $\boldsymbol{\varepsilon}_t$  in (2.1) consist of independent  $N(0, 1)$  components and they are also independent across  $t$ . We set  $k_0 = 1$  in (2.6) and (2.7). For each setting, we replicate the simulation 200 times.

Table 1 reports the relative frequency estimates for the probability  $P(\hat{r} = r) = P(\hat{r} = 3)$  with  $\delta = 0$  and  $0.5$ . The estimation performs better when the factors are stronger. Even when the factors are weak (i.e.,  $\delta = 0.5$ ), the estimation for  $r$  is very accurate for  $n \geq 800$ . When the factors are strong (i.e.,  $\delta = 0$ ), we observe a phenomenon coined as “blessing of dimensionality” in the sense that the estimation for  $r$  improves as the dimension  $p$  increases. For example, when the sample size  $n = 100$ , the relative frequencies for  $\hat{r} = r$  are, respectively, 0.68, 0.8, 0.815 and 0.82 for  $p = 20, 50, 80$  and  $120$ . The improvement is due to the increased information on  $r$  from the added

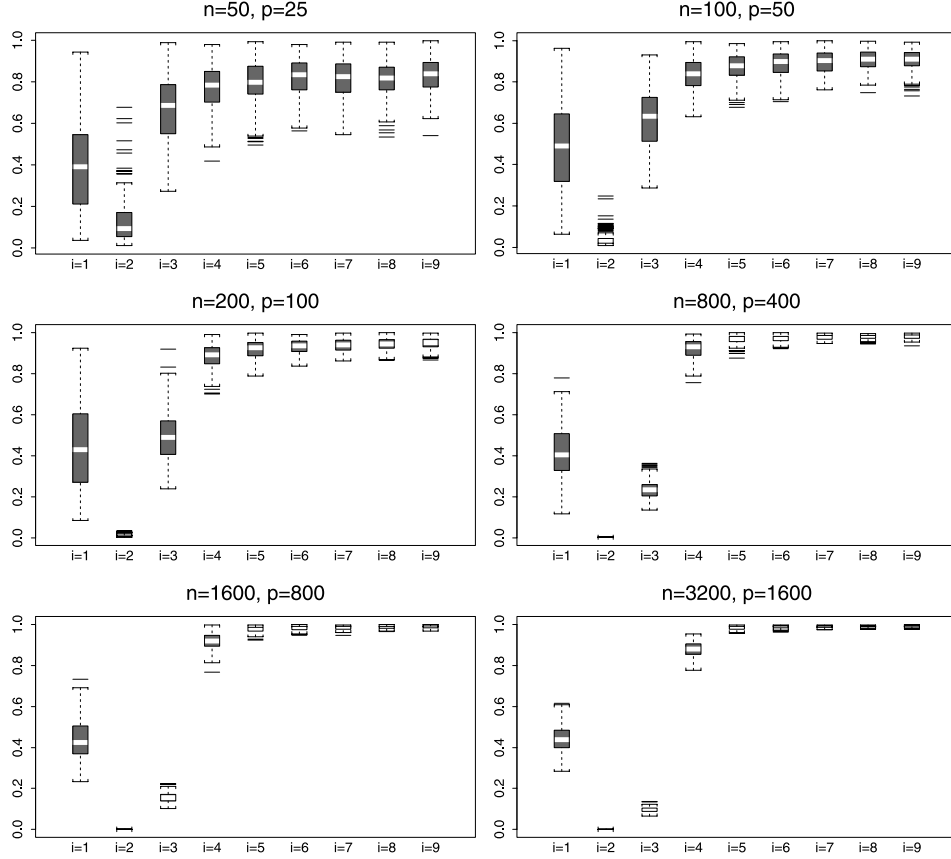


FIG. 3. *Boxplots for the ratios  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$  with two strong factors ( $\delta = 0$ ) and one weak factor ( $\delta = 0.5$ ) and  $r = 3$ ,  $p = n/2$ .*

components of  $\mathbf{y}_t$  when  $p$  increases. When  $\delta = 0.5$ , the columns of  $\mathbf{A}$  are  $p$ -vectors with the norm  $p^{0.25}$  [see (3.1)]. Hence we may think that many elements of  $\mathbf{A}$  are now effectively 0. The increase of the information on the factors is coupled with the increase of “noise” when  $p$  increases. Indeed, Table 1 shows that when factors are weak as  $\delta = 0.5$ , the estimation for  $r$  does not necessarily improve as  $p$  increases.

We also experiment with a setting with two strong factors (with  $\delta = 0$ ) and one weak factor (with  $\delta = 0.5$ ). Then the ratio-based estimator  $\hat{r}$  tends to take two values, picking up the two strong factors only. However Figure 3 indicates that the information on the third weak factor is not lost. In fact,  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$  tends to take the second smallest value at  $i = 3$ . In this case a two-step estimation procedure should be employed in order to identify the number of factors correctly; see Section 4 below.

3.4. *Improved rates for the estimated eigenvalues.* The rates in Theorem 1 can be further improved, if we are prepared to entertain some additional conditions on  $\varepsilon_t$  in model (2.1). Such an improvement is relevant as the condition that  $h_n = p^\delta n^{-1/2} \rightarrow 0$ , required in Theorem 1, is sometimes unnecessary. For example, in Table 1, the ratio-based estimator  $\hat{r}$  works perfectly well when  $\delta = 0.5$  and  $n$  is sufficiently large (e.g.,  $n \geq 800$ ), even though  $h_n = (p/n)^{1/2} \not\rightarrow 0$ . Furthermore, in relation to the phenomenon of “blessing of dimensionality” exhibited in Table 1, Theorem 1 fails to reflect the possible improvement on the estimation for  $r$  when  $p$  increases; see also Remark 2(i). We first introduce some additional conditions on  $\varepsilon_t$ :

(C7) Let  $\varepsilon_{jt}$  denote the  $j$ th component of  $\varepsilon_t$ . Then  $\varepsilon_{jt}$  are independent for different  $t$  and  $j$ , and have mean 0 and common variance  $\sigma^2 < \infty$ .

(C8) The distribution of each  $\varepsilon_{jt}$  is symmetric. Furthermore,  $E(\varepsilon_{jt}^{2k+1}) = 0$ , and  $E(\varepsilon_{jt}^{2k}) \leq (\tau k)^k$  for all  $1 \leq j \leq p$  and  $t, k \geq 1$ , where  $\tau > 0$  is a constant independent of  $j, t, k$ .

(C9) All the eigenvalues of  $\Sigma_\varepsilon$  are uniformly bounded as  $p \rightarrow \infty$ .

The moment condition  $E(\varepsilon_{jt}^{2k}) \leq (\tau k)^k$  in (C8) implies that  $\varepsilon_{jt}$  are sub-Gaussian. Condition (C9) imposes some constraint on the correlations among the components of  $\varepsilon_t$ . When all components of  $\{\varepsilon_t\}$  are independent  $N(0, \sigma^2)$ , (C7)–(C9) hold. See also conditions (i')–(iv') of [17].

**THEOREM 2.** *Let conditions (C1)–(C8) hold,  $\ell_n \equiv p^{\delta/2} n^{-1/2} \rightarrow 0$  and  $n = O(p)$ . Then as  $p, n \rightarrow \infty$ , the following assertions hold:*

- (i)  $|\hat{\lambda}_j - \lambda_j| = O_P(p^{2-3\delta/2} n^{-1/2})$  for  $j = 1, \dots, r$ ,
- (ii)  $\hat{\lambda}_j = O_P(p^{2-\delta} n^{-1})$  for  $j = r+1, \dots, (k_0+1)r$ ,
- (iii)  $\hat{\lambda}_j = O_P(p^2 n^{-2})$  for  $j = (k_0+1)r+1, \dots, p$ .

*If in addition (C9) holds, the rate in (ii) above can be further improved to*

$$(3.4) \quad \hat{\lambda}_j = O_P(p^{3/2-\delta} n^{-1/2}), \quad j = r+1, \dots, (k_0+1)r.$$

**COROLLARY 2.** *Under the conditions of Theorem 2, it holds that*

$$\hat{\lambda}_{j+1}/\hat{\lambda}_j \asymp 1, \quad j = 1, \dots, r-1, \quad \text{and} \quad \hat{\lambda}_{r+1}/\hat{\lambda}_r = O_P(p^\delta n^{-1}).$$

*If in addition (C9) also holds,  $\hat{\lambda}_{r+1}/\hat{\lambda}_r = O_P(p^{\delta-1/2} n^{-1/2})$ .*

The proofs of Theorem 2 and Corollary 1 are given in the [Appendix](#).

**REMARK 3.** (i) By comparing with Theorem 1, the error rate for nonzero  $\lambda_j$  in Theorem 2 is improved by a factor  $p^{-\delta/2}$ , the error rate for zero-eigenvalues is by a factor  $p^{-\delta}$  at least. However, those estimators themselves may still diverge, as illustrated in Figure 1.

(ii) Theorem 2(iii) is an interesting consequence of the random matrix theory. The key message here is as follows: for the eigenvalues corresponding purely to the matrix  $\sum_{k=1}^{k_0} \widehat{\Sigma}_\varepsilon(k) \widehat{\Sigma}_\varepsilon(k)'$ , their magnitudes adjusted for  $p^{2-2\delta}$  converge at a super-fast rate. The matrix  $\sum_{k=1}^{k_0} \widehat{\Sigma}_\varepsilon(k) \widehat{\Sigma}_\varepsilon(k)'$  is a part of  $\widehat{\mathbf{M}}$  in (2.7), where  $\widehat{\Sigma}_\varepsilon(k)$  is the sample lag- $k$  autocovariance matrix for  $\{\varepsilon_t\}$ . In particular, when all the factors are strong (i.e.,  $\delta = 0$ ), the convergence rate is  $n^{-2}$ . Such a super convergence rate never occurs when  $p$  is fixed.

(iii) Condition  $\ell_n \rightarrow 0$  is mild, and is weaker than condition  $h_n \rightarrow 0$  required in Theorem 1. For example, when  $p \asymp n$ , this condition is implied by the condition  $\delta \in [0, 1)$ .

(iv) With additional condition (C9),  $\widehat{\lambda}_{r+1}/\widehat{\lambda}_r = O_P(p^{-1/2}n^{-1/n})$  when all factors are strong. This shows that the speed at which  $\widehat{\lambda}_{r+1}/\widehat{\lambda}_r$  converges to 0 increases when  $p$  increases. This property gives a theoretical explanation why the identification for  $r$  becomes easier for larger  $p$  when all factors are strong (i.e.,  $\delta = 0$ ). See Table 1.

**4. Two-step estimation.** In this section, we outline a two-step estimation procedure. We will show that it is superior than the one-step procedure presented in Section 2.2 for the determination of the number of factors as well as for the estimation of the factor loading matrices in the presence of the factors with different degrees of strength. A similar procedure is described in [19] to improve the estimation for factor loading matrices in the presence of small eigenvalues, although they gave no theoretical underpinning on why and when such a procedure is advantageous.

Consider model (2.1) with  $r_1$  strong factors with strength  $\delta_1 = 0$  and  $r_2$  weak factors with strength  $\delta_2 > 0$ , where  $r_1 + r_2 = r$ . Now (2.1) may be written as

$$(4.1) \quad \mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \varepsilon_t = \mathbf{A}_1\mathbf{x}_{1t} + \mathbf{A}_2\mathbf{x}_{2t} + \varepsilon_t,$$

where  $\mathbf{x}_t = (\mathbf{x}'_{1t} \ \mathbf{x}'_{2t})'$ ,  $\mathbf{A} = (\mathbf{A}_1 \ \mathbf{A}_2)$  with  $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$ ,  $\mathbf{x}_{1t}$  consists of  $r_1$  strong factors, and  $\mathbf{x}_{2t}$  consists of  $r_2$  weak factors. Like model (2.1) in Section 2.1,  $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$  and  $\mathbf{x}_t = (\mathbf{x}_{1t}, \mathbf{x}_{2t})$  are not uniquely defined, but only  $\mathcal{M}(\mathbf{A})$  is. Hereafter  $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$  corresponds to a suitably rotated version of the original  $\mathbf{A}$  in model (4.1), where now  $\mathbf{A}$  contains all the eigenvectors of  $\mathbf{M}$  corresponding to its nonzero eigenvalues. Refer to (2.6) for the definition of  $\mathbf{M}$ .

To present the two-step estimation procedure clearly, let us assume that we know  $r_1$  and  $r_2$  first. Using the method in Section 2.2, we first obtain the estimator  $\widehat{\mathbf{A}} \equiv (\widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2)$  for the factor loading matrix  $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$ , where the columns of  $\widehat{\mathbf{A}}_1$  are the  $r_1$  orthonormal eigenvectors of  $\widehat{\mathbf{M}}$  corresponding to its  $r_1$  largest eigenvalues. In practice we may identify  $r_1$  using, for example, the ratio-based estimation method (2.8); see Figure 3. We carry out the



second-step estimation as follows. Let

$$(4.2) \quad \mathbf{y}_t^* = \mathbf{y}_t - \widehat{\mathbf{A}}_1 \widehat{\mathbf{A}}_1' \mathbf{y}_t$$

for all  $t$ . We perform the same estimation for data  $\{\mathbf{y}_t^*\}$  now, and obtain the  $p \times r_2$  estimated factor loading matrix  $\widetilde{\mathbf{A}}_2$  for the  $r_2$  weak factors. Combining the two estimators together, we obtain the final estimator for  $\mathbf{A}$  as

$$(4.3) \quad \widetilde{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \widetilde{\mathbf{A}}_2).$$

Theorem 3 below presents the convergence rates for both the one-step estimator  $\widehat{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2)$  and the two-step estimator  $\widetilde{\mathbf{A}} = (\widehat{\mathbf{A}}_1, \widetilde{\mathbf{A}}_2)$ . It shows that  $\widetilde{\mathbf{A}}$  converges to  $\mathbf{A}$  at a faster rate than  $\widehat{\mathbf{A}}$ . The results are established with known  $r_1$  and  $r_2$ . In practice we estimate  $r_1$  and  $r_2$  using the ratio-based estimators. See also Theorem 4 below. We introduce some regularity conditions first. Let  $\boldsymbol{\Sigma}_{12}(k) = \text{Cov}(\mathbf{x}_{1,t+k}, \mathbf{x}_{2t})$ ,  $\boldsymbol{\Sigma}_{21}(k) = \text{Cov}(\mathbf{x}_{2,t+k}, \mathbf{x}_{1t})$ ,  $\boldsymbol{\Sigma}_i(k) = \text{Cov}(\mathbf{x}_{i,t+k}, \mathbf{x}_{it})$  and  $\boldsymbol{\Sigma}_{i\epsilon}(k) = \text{Cov}(\mathbf{x}_{i,t+k}, \boldsymbol{\varepsilon}_t)$  for  $i = 1, 2$ :

(C5)' For  $i = 1, 2$ ,  $1 \leq k \leq k_0$ ,  $\|\boldsymbol{\Sigma}_i(k)\| \asymp p^{1-\delta_i} \asymp \|\boldsymbol{\Sigma}_i(k)\|_{\min}$ ,  $\|\boldsymbol{\Sigma}_{21}(k)\| \asymp \|\boldsymbol{\Sigma}_{21}(k)\|_{\min}$  and  $\|\boldsymbol{\Sigma}_{12}(k)\| = O(p^{1-\delta_2/2})$ .

(C6)'  $\text{Cov}(\mathbf{x}_t, \boldsymbol{\varepsilon}_s) = 0$  for any  $t, s$ .

The condition on  $\boldsymbol{\Sigma}_i(k)$  in (C5)' is an analogue to condition (C5). See Remark 1(i) in Section 3.2 for the background of those conditions. The order of  $\|\boldsymbol{\Sigma}_{21}(k)\|_{\min}$  will be specified in the theorems below. The order of  $\|\boldsymbol{\Sigma}_{12}(k)\|$  is not restrictive, since  $p^{1-\delta_2/2}$  is the largest possible order when  $\delta_1 = 0$ . See also the discussion in Remark 1(ii). Condition (C6)' replaces condition (C6). Here we impose a strong condition  $\boldsymbol{\Sigma}_{i\epsilon}(k) = 0$  to highlight the benefits of the two-step estimation procedure. See Remark 4(iii) below. Put

$$\mathbf{W}_i = (\boldsymbol{\Sigma}_i(1), \dots, \boldsymbol{\Sigma}_i(k_0)), \quad \mathbf{W}_{21} = (\boldsymbol{\Sigma}_{21}(1), \dots, \boldsymbol{\Sigma}_{21}(k_0)).$$

**THEOREM 3.** *Let conditions (C1)–(C4), (C5)', (C6)', (C7) and (C8) hold. Let  $n = O(p)$  and  $\kappa_n \equiv p^{\delta_2/2} n^{-1/2} \rightarrow 0$ , as  $n \rightarrow \infty$ . Then it holds that*

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\| = O_P(n^{-1/2}), \quad \|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(\kappa_n) = \|\widetilde{\mathbf{A}} - \mathbf{A}\|.$$

Furthermore,

$$\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(\nu_n) = \|\widehat{\mathbf{A}} - \mathbf{A}\|,$$

if, in addition,  $\nu_n \rightarrow 0$  and  $p^{c\delta_2} n^{-1/2} \rightarrow 0$ , where  $\nu_n$  and  $c$  are defined as follows:

$$\nu_n = \begin{cases} p^{\delta_2} \kappa_n, & \text{if } \|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2}) \text{ } (c = 1); \\ p^{(2c-1)\delta_2} \kappa_n, & \text{if } \|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2} \text{ for } 1/2 \leq c < 1, \text{ and} \\ & \|\mathbf{W}_1 \mathbf{W}_{21}'\| \leq q \|\mathbf{W}_1\|_{\min} \|\mathbf{W}_{21}\| \text{ for } 0 \leq q < 1. \end{cases}$$

Note that  $\kappa_n/\nu_n \rightarrow 0$ . Theorem 3 indicates that between  $\mathbf{A}_1$  and  $\mathbf{A}_2$ , the latter is more difficult to estimate, and the convergence rate of an estimator for  $\mathbf{A}$  is determined by the rate for  $\mathbf{A}_2$ . This is intuitively understandable

as the coefficient vectors for weak factors effectively contain many zero-components; see (3.1). Therefore a nontrivial proportion of the components of  $\mathbf{y}_t$  may contain little information on weak factors. When  $\|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}$ ,  $\|\mathbf{W}_2\|$  is dominated by  $\|\mathbf{W}_{21}\|_{\min}$ . The condition  $\|\mathbf{W}_1\mathbf{W}'_{21}\| \leq q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\|$  for  $0 \leq q < 1$  is imposed to control the behavior of the  $(r_1 + 1)$ th to the  $r$ th largest eigenvalues of  $\mathbf{M}$  under this situation. If this is not valid, those eigenvalues can become very small and give a bad estimator for  $\mathbf{A}_2$ , and thus  $\mathbf{A}$ . Under this condition, the structure of the autocovariance for the strong factors, and the structure of the cross-autocovariance between the strong and weak factors, are not similar.

Recall that  $\lambda_j$  and  $\hat{\lambda}_j$  are the  $j$ th largest eigenvalue of, respectively,  $\mathbf{M}$  defined in (2.6) and  $\widehat{\mathbf{M}}$  defined in (2.7). We define matrices  $\mathbf{M}^*$  and  $\widehat{\mathbf{M}}^*$  in the same manner as  $\mathbf{M}$  and  $\widehat{\mathbf{M}}$  but with  $\{\mathbf{y}_t\}$  replaced by  $\{\mathbf{y}_t^*\}$  defined in (4.2), and denote by  $\lambda_j^*$  and  $\hat{\lambda}_j^*$  the  $j$ th largest eigenvalue of, respectively,  $\mathbf{M}^*$  and  $\widehat{\mathbf{M}}^*$ . The following theorem shows the different behavior of the ratio of eigenvalues under the one-step and two-step estimation. Readers who are interested in the explicit rates for the eigenvalues are referred to Lemma 1 in the Appendix.

**THEOREM 4.** *Under the same conditions of Theorem 3, the following assertions hold:*

- (i) For  $1 \leq i < r_1$  or  $r_1 < i < r$ ,  $\hat{\lambda}_{i+1}/\hat{\lambda}_i \asymp 1$ . For  $1 \leq j < r_2$ ,  $\hat{\lambda}_{j+1}^*/\hat{\lambda}_j^* \asymp 1$ .
- (ii)  $\hat{\lambda}_{r+1}/\hat{\lambda}_r \xrightarrow{P} 0$  and  $\hat{\lambda}_{r_1+1}/\hat{\lambda}_{r_1} = o_p(\hat{\lambda}_{r+1}/\hat{\lambda}_r)$  provided  $\delta_2 > 1/(8c-1)$ ,  $p^{(1-\delta_2)/2}n^{-1/2} \rightarrow 0$ ,  $p^{(6c-1/2)\delta_2-1/2}n^{-1/2} \rightarrow \infty$ .
- (iii)  $\hat{\lambda}_{r+1}/\hat{\lambda}_r \xrightarrow{P} 0$  and  $\hat{\lambda}_{r_2+1}^*/\hat{\lambda}_{r_2}^* = o_p(\hat{\lambda}_{r+1}/\hat{\lambda}_r)$  provided  $p^{(4c-3/2)\delta_2-1/2} \times n^{1/2} \rightarrow \infty$ .

**REMARK 4.** (i) Theorem 4(i) and (ii) imply that the one-step estimation is likely to lead to  $\hat{r} = r_1$ . For instance, when  $p \asymp n$ , then Theorem 4(ii) says that  $\hat{\lambda}_{r_1+1}/\hat{\lambda}_{r_1}$  has a faster rate of convergence than  $\hat{\lambda}_{r+1}/\hat{\lambda}_r$  as long as  $\delta_2 > 2/5$ . Figure 3 shows exactly this situation.

(ii) Theorem 4(iii) implies that the two-step estimation is more capable to identify the additional  $r_2$  factors than the one-step estimation. In particular, if  $p \asymp n$ ,  $\hat{\lambda}_{r_2+1}^*/\hat{\lambda}_{r_2}^*$  always has a faster rate of convergence than  $\hat{\lambda}_{r+1}/\hat{\lambda}_r$ . Unfortunately we are unable to establish the asymptotic properties for  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$  for  $i > r$ , and  $\hat{\lambda}_{j+1}^*/\hat{\lambda}_j^*$  for  $j > r_2$ , though we believe that conjectures similar to (3.3) continue to hold.

(iii) When  $\delta_1 > 0$  and/or the cross-autocovariances between different factors and the noise are stronger, the similar and more complex results can be established via more involved algebra in the proofs.

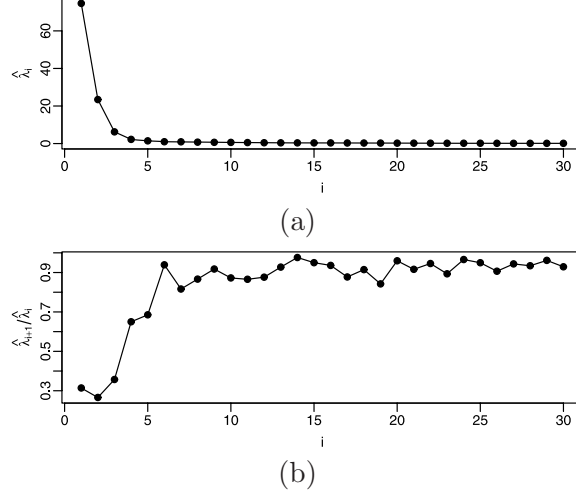


FIG. 4. Plots of the estimated eigenvalues (a) and the ratios of estimated eigenvalues of  $\widehat{\mathbf{M}}$  (b) for Example 1.

**5. Real data examples.** We illustrate our method using two real data sets.

**EXAMPLE 1.** We first analyze the daily returns of 123 stocks in the period 2 January 2002–11 July 2008. Those stocks were selected among those included in the S&P500 and were traded every day during the period. The returns were calculated in percentages based on the daily close prices. We have in total  $n = 1642$  observations with  $p = 123$ . We apply the eigenanalysis to the matrix  $\widehat{\mathbf{M}}$  defined in (2.7) with  $k_0 = 5$ . The obtained eigenvalues (in descending order) and their ratios are plotted in Figure 4. It is clear that the ratio-based estimator (2.8) leads to  $\hat{r} = 2$ , indicating two factors. Varying the value of  $k_0$  between 1 and 100 in the definition of  $\widehat{\mathbf{M}}$  leads to little change in the ratios  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$ , and the estimate  $\hat{r} = 2$  remains unchanged. Figure 4(a) shows that  $\hat{\lambda}_i$  is close to 0 for all  $i \geq 5$ . Figure 4(b) indicates that the ratio  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$  is close to 1 for all large  $i$ , which is in line with conjecture (3.3).

The first two panels of Figure 5 display the time series plots of the two component series of the estimated factors  $\hat{\mathbf{x}}_t$  defined as in (2.2). Their cross-autocorrelations are presented in Figure 6. Although each of the two estimated factors shows little significant autocorrelation, there are some significant cross-correlations between the two series. The cross-autocorrelations of the three residual series  $\hat{\gamma}'_j \mathbf{y}_t$  for  $j = 3, 4, 5$  are not significantly different from 0, where  $\hat{\gamma}_j$  is the unit eigenvector of  $\widehat{\mathbf{M}}$  corresponding to its  $j$ th largest eigenvalue. If there were any serial correlations left in the data after extracting the two estimated factors, those correlations are most likely to show up in those three residual series.

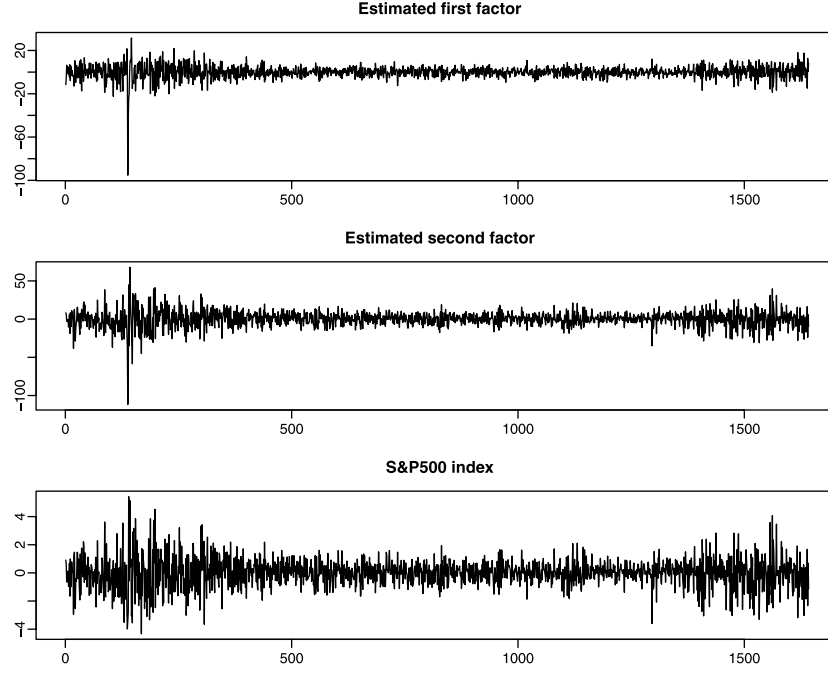


FIG. 5. The time series plots of the two estimated factors and the return series of the S&P500 index in the same time period.

Figure 4 may suggest the existence of a third and weaker factor, though there are hardly any significant autocorrelations in the series  $\hat{\gamma}_3 \mathbf{y}_t$ . In fact  $\hat{\lambda}_3 = 6.231$  and  $\hat{\lambda}_4/\hat{\lambda}_3 = 0.357$ . Note that now  $\hat{\lambda}_j$  is not necessarily a consistent estimator for  $\lambda_j$  although  $\hat{\lambda}_{r+1}/\hat{\lambda}_r \xrightarrow{P} 0$ ; see Theorem 1(ii) and Corollary 1. To investigate this further, we apply the two-step estimation procedure presented in Section 4. By subtracting the two estimated factors from the above, we obtain the new data  $\mathbf{y}_t^*$  [see (4.3)]. We then calculate the eigenvalues and their ratios of the matrix  $\widehat{\mathbf{M}}^*$ . The minimum value of the ratios is  $\hat{\lambda}_2^*/\hat{\lambda}_1^* = 0.667$ , which is closely followed by  $\hat{\lambda}_3^*/\hat{\lambda}_2^* = 0.679$  and  $\hat{\lambda}_4^*/\hat{\lambda}_3^* = 0.744$ . There is no evidence to suggest that  $\hat{\lambda}_2^*/\hat{\lambda}_1^* \rightarrow 0$ ; see Theorem 4. This reinforces our choice  $\hat{r} = 2$ .

With  $p$  as large as 123, it is difficult to gain insightful interpretation on the estimated factors by looking through the coefficients in  $\widehat{\mathbf{A}}$  [see (2.2)]. To link our fitted factor model with some classical asset pricing theory in finance, we wonder if the market index (i.e., the S&P500 index) is a factor in our fitted model, or more precisely, if it can be written as a linear combination of the two estimated factors. When this is true,  $\mathbf{P}\mathbf{u} = 0$ , where  $\mathbf{u}$  is the  $1642 \times 1$  vector consisting of the returns of the S&P500 index over the same time period, and  $\mathbf{P}$  denotes the projection matrix onto the orthogonal

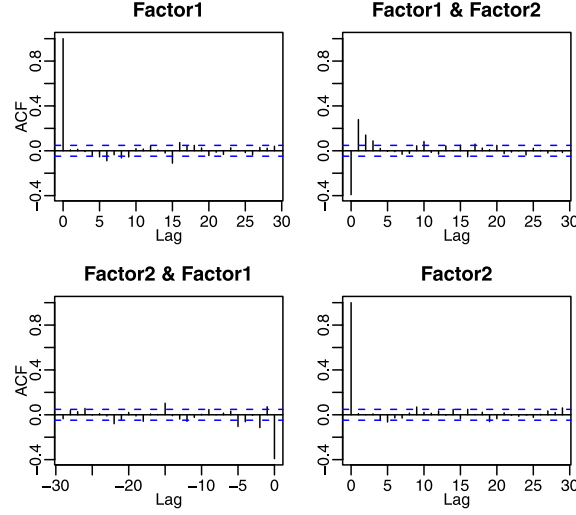


FIG. 6. The cross-autocorrelations of the two estimated factors for Example 1.

complement of the linear space spanned by the two component series  $\hat{\mathbf{x}}_t$ , which is a 1640-dimensional subspace in  $R^{1642}$ . This S&P500 return series is plotted together with the two component series  $\hat{\mathbf{x}}_t$  in Figure 5. It turns out that  $\|\mathbf{P}\mathbf{u}\|^2$  is not exactly 0 but  $\|\mathbf{P}\mathbf{u}\|^2/\|\mathbf{u}\|^2 = 0.023$ , that is, the 97.7% of the S&P500 returns can be expressed as a linear combination of the two estimated factors. Thus our analysis suggests the following model for  $\mathbf{y}_t$ —the daily returns of the 123 stocks:

$$\mathbf{y}_t = \mathbf{a}_1 u_t + \mathbf{a}_2 v_t + \boldsymbol{\varepsilon}_t,$$

where  $u_t$  denotes the return of the S&P500 on the day  $t$ ,  $v_t$  is another factor, and  $\boldsymbol{\varepsilon}_t$  is a  $123 \times 1$  vector white-noise process.

Figure 5 shows that there is an early period with big sparks in the two estimated factor processes. Those sparks occurred around 24 September 2002 when the markets were highly volatile and the Dow Jones Industrial Average had lost 27% of the value it held on 1 January 2001. However, those sparks are significantly less extreme in the returns of the S&P500 index; see the third panel in Figure 5. In fact the projected S&P500 return  $\mathbf{P}\mathbf{u}$  is the linear combination of those two estimated factors with the coefficients  $(-0.0548, 0.0808)$ . Two observations may be drawn from the opposite signs of those two coefficients: (i) there is an indication that those two factors draw the energy from the markets with opposite directions, and (ii) the portfolio S&P500 index hedges the risks across different markets.

EXAMPLE 2. We analyze a set of monthly average sea surface air pressure records (in Pascal) from January 1958 to December 2001 (i.e., 528 months in total) over a  $10 \times 44$  grid in a range of  $22.5^\circ$ – $110^\circ$  longitude in the

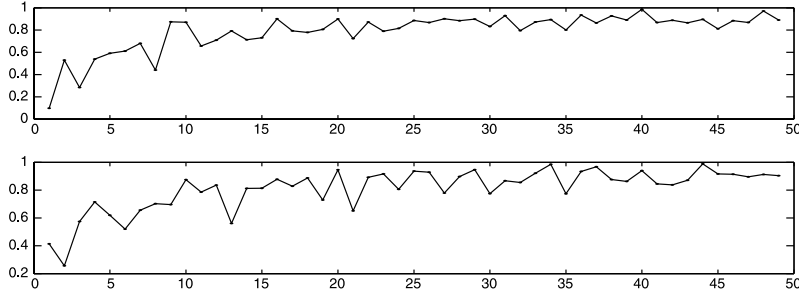


FIG. 7. Plots of  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$ —the ratio of the eigenvalues of  $\widehat{\mathbf{M}}$  (the top panel) and  $\widehat{\mathbf{M}}^*$  (the bottom panel), against  $i$ , for Example 2.

North Atlantic Ocean. Let  $P_t(u, v)$  denote the air pressure in the  $t$ th month at the location  $(u, v)$ , where  $u = 1, \dots, 10, v = 1, \dots, 44$  and  $t = 1, \dots, 528$ . We first subtract each data point by the monthly mean over the 44 years at its location:  $\frac{1}{44} \sum_{i=1}^{44} P_{12(i-1)+j}(u, v)$ , where  $j = 1, \dots, 12$ , representing the 12 different months over a year. We then line up the new data over  $10 \times 44 = 440$  grid points as a vector  $\mathbf{y}_t$ , so that  $\mathbf{y}_t$  is a  $p$ -variate time series with  $p = 440$ . We have  $n = 528$  observations.

To fit the factor model (2.1) to  $\mathbf{y}_t$ , we calculate the eigenvalues and the eigenvectors of the matrix  $\widehat{\mathbf{M}}$  defined in (2.7) with  $k_0 = 5$ . Let  $\hat{\lambda}_1 > \hat{\lambda}_2 > \dots$  denote the eigenvalues of  $\widehat{\mathbf{M}}$ . The ratios  $\hat{\lambda}_{i+1}/\hat{\lambda}_i$  are plotted against  $i$  in the top panel of Figure 7 which indicates the ratio-based estimate for the number of factor is  $\hat{r} = 1$ ; see (2.8). However, the second smallest ratio is  $\hat{\lambda}_4/\hat{\lambda}_3$ . This suggests that there may exist two weaker factors in addition; see Theorem 4(ii) and also Figure 3. We adopt the two-step estimation procedure presented in Section 4 to identify the factors of different strength. By removing the factor corresponding to the largest eigenvalue of  $\widehat{\mathbf{M}}$ , the resulting “residuals” are denoted as  $\mathbf{y}_t^*$ ; see (4.2). Now we repeat the factor modeling for data  $\mathbf{y}_t^*$ , and plot the ratios of eigenvalues of matrix  $\widehat{\mathbf{M}}^*$  in the second panel of Figure 7. It shows clearly the minimum value at 2, indicating further two (weaker) factors. Combining the above two steps together, we set  $\hat{r} = 3$  in the fitted model. We repeated the above calculation with  $k_0 = 1$  in (2.7). We still find three factors with the two-step procedure, and the estimated factors series are very similar to the case when  $k_0 = 5$ . This is consistent with the simulation results in [13], where they showed empirically that the estimated factor models are not sensitive to the choice of  $k_0$ .

We present the time series plots for the three estimated factors  $\tilde{\mathbf{x}}_t = \tilde{\mathbf{A}}' \mathbf{y}_t$  in Figure 8, where  $\tilde{\mathbf{A}}$  is a  $440 \times 3$  matrix with the first column being the unit eigenvector of  $\widehat{\mathbf{M}}$  corresponding to its largest eigenvalue, and the other two columns being the orthonormal eigenvectors of  $\widehat{\mathbf{M}}^*$  corresponding to its two largest eigenvalues; see (4.3) and also (2.2). They collectively account

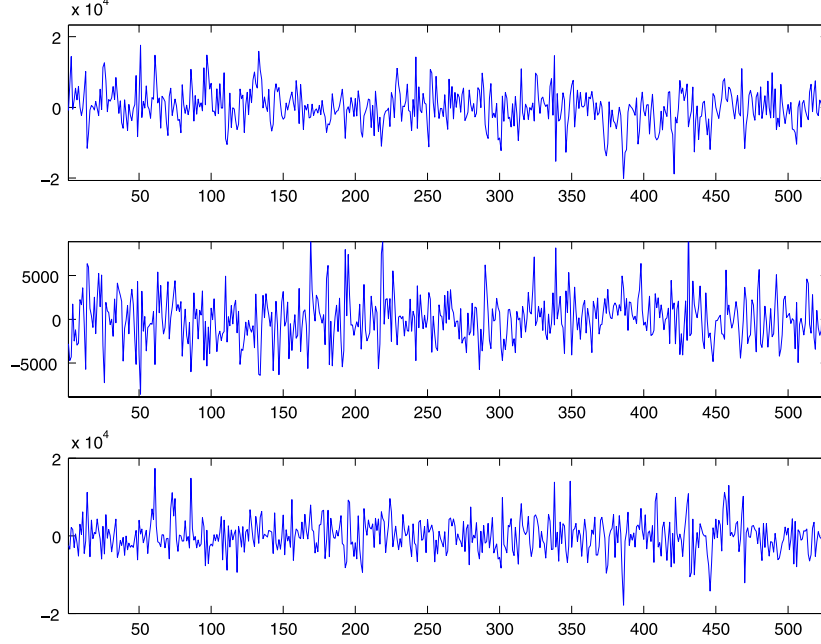


FIG. 8. Time series plot of the three estimated factors for Example 2.

for 85.3% of the total variation in  $\mathbf{y}_t$  which has 440 components. In fact each of the three factors accounts for, respectively, 57.5%, 18.2% and 9.7% of the total variation of  $\mathbf{y}_t$ . Figure 9 depicts the factor loading surfaces of the three factors. Some interesting regional patterns are observed from those plots. For example, the first factor is the main driving force for the dynamics in the north and especially the northeast. The second factor influences the dynamics in the east and the west in the opposite directions, and has little impact in the narrow void between them. The third factor impacts mainly the dynamics of the southeast region. We also notice that none of those factors can be seen as idiosyncratic components as each of them affects quite a large number of locations.

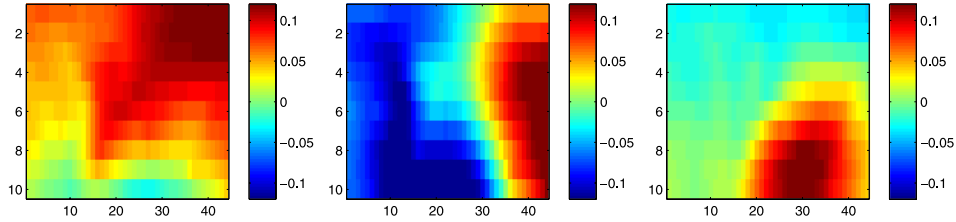


FIG. 9. Factor loading surface of the first, second and third factors (from left to right) for Example 2.



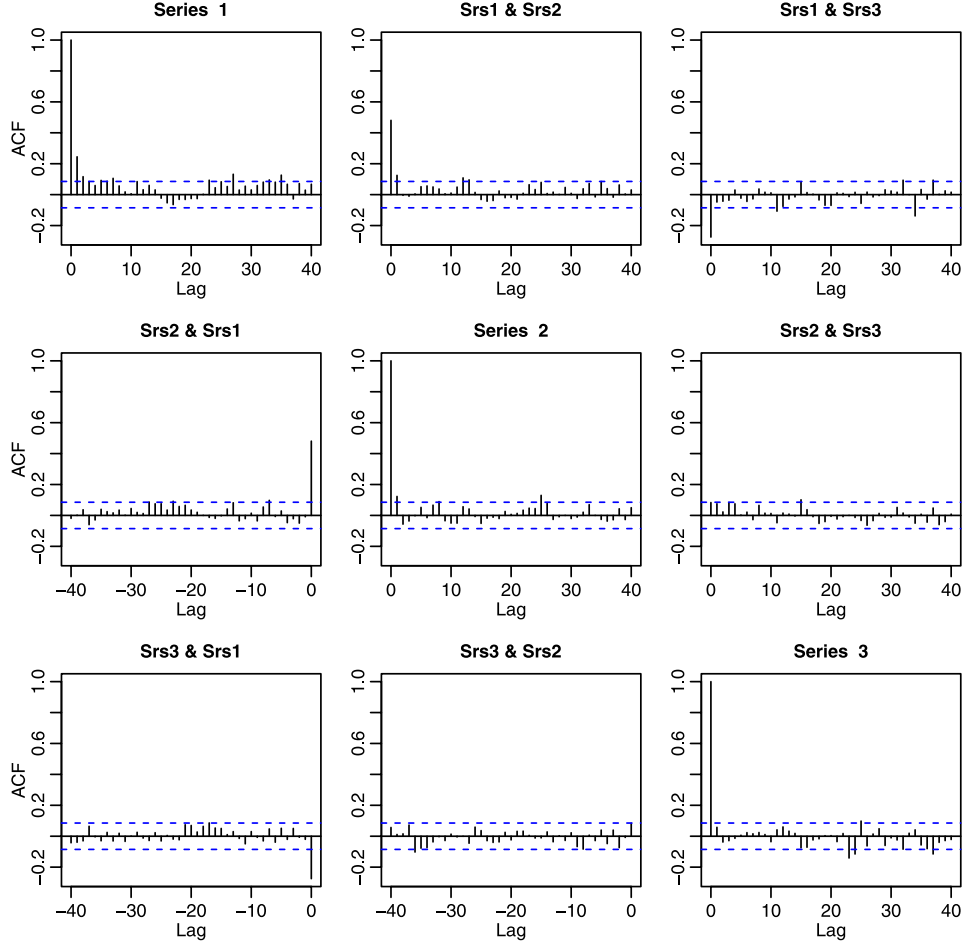


FIG. 10. *Example 2: sample cross-correlation functions for the three estimated factors.*

Figure 10 presents the sample cross-correlations for the three estimated factors. It shows significant, though small, autocorrelations or cross-correlations at some nonzero lags. Figure 11 is the sample cross-correlations for three residuals series selected from three locations for which one is far apart from the other two spatially, showing little autocorrelations at nonzero lags. This indicates that our approach is capable to identify the factors based on serial correlations.

Finally we note that the BIC method of [2] yields the estimate  $\hat{r} = n = 528$  for this particular data set. We suspect that this may be due to the fact that [2] requires all the eigenvalues of  $\Sigma_\varepsilon$  be uniformly bounded when  $p \rightarrow \infty$ . This may not be the case for this particular data set, as the nearby locations

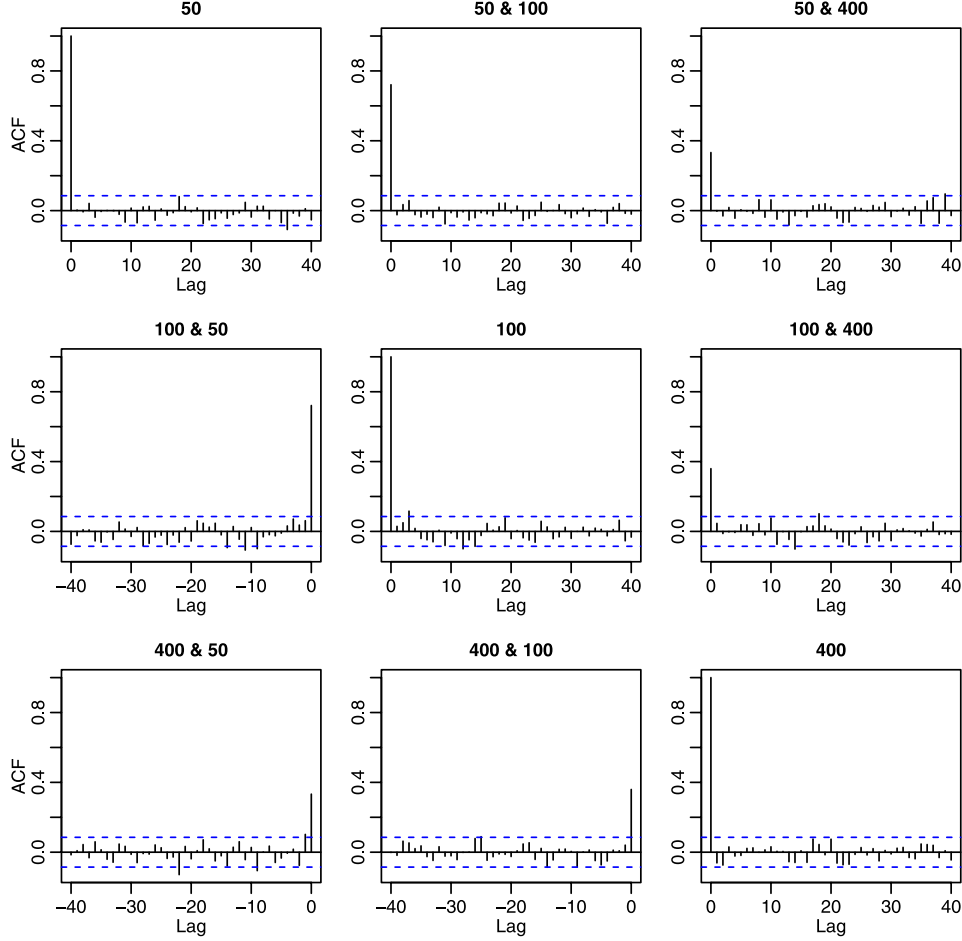


FIG. 11. *Example 2: sample cross-correlation functions for three residual series. 50 represents grid position (10, 5), 100 for (10, 10) and 400 for (10, 40).*

are strongly spatially correlated, which may lead to very large and also very small eigenvalues for  $\Sigma_\varepsilon$ . Indeed, for this data set, the three largest eigenvalues of  $\hat{\Sigma}_\varepsilon$  are on the order of  $10^6$ , and the three smallest eigenvalues are practically 0. Since the typical magnitude of  $\hat{\varepsilon}_t$  is  $10^2$  from our analysis, we have done simulations (not shown here) showing that the typical largest eigenvalues for  $\hat{\Sigma}_\varepsilon$ , if  $\{\varepsilon_t\}$  is weakly correlated white noise, should be around  $10^4$  to  $10^5$ , and the smallest around  $10^2$  to  $10^3$  when  $p = 440$  and  $n = 528$ . Such a huge difference in the magnitude of the eigenvalues suggests strongly that the components of the white-noise vector  $\varepsilon_t$  are strongly correlated. Our method does not require the uniform boundedness of the eigenvalues of  $\Sigma_\varepsilon$ .

## APPENDIX

PROOF OF THEOREM 1. We present some notational definitions first. We denote by  $\hat{\lambda}_j, \hat{\gamma}_j$  the  $j$ th largest eigenvalue of  $\widehat{\mathbf{M}}$  and the corresponding orthonormal eigenvector, respectively. The corresponding population values are denoted by  $\lambda_j$  and  $\mathbf{a}_j$  for the matrix  $\mathbf{M}$ . Hence  $\widehat{\mathbf{A}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_r)$  and  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_r)$ . We also have

$$\lambda_j = \mathbf{a}_j' \mathbf{M} \mathbf{a}_j, \quad \hat{\lambda}_j = \hat{\gamma}_j' \widehat{\mathbf{M}} \hat{\gamma}_j, \quad j = 1, \dots, p.$$

We show some intermediate results now. With conditions (C3) and (C5) and the fact that  $\{\varepsilon_t\}$  is white noise, we have

$$(A.1) \quad \begin{aligned} \|\widehat{\Sigma}_x(k) - \Sigma_x(k)\| &= O_P(p^{1-\delta} n^{-1/2}), \\ \|\widehat{\Sigma}_{x\varepsilon}(k) - \Sigma_{x\varepsilon}(k)\|, \quad \|\widehat{\Sigma}_{\varepsilon x}(k)\| &= O_P(p^{1-\delta/2} n^{-1/2}), \end{aligned}$$

where  $k = 0, 1, \dots, k_0$ . Then following the proof of Theorem 1 of [13], we have the following for  $k = 1, \dots, k_0$ :

$$(A.2) \quad \begin{aligned} \|\widehat{\mathbf{M}} - \mathbf{M}\| &= O_P(\|\Sigma_y(k)\| \cdot \|\widehat{\Sigma}_y(k) - \Sigma_y(k)\|) \\ \text{where } \|\Sigma_y(k)\| &= O(p^{1-\delta}), \\ \|\widehat{\Sigma}_y(k) - \Sigma_y(k)\| &= O_P(p^{1-\delta} n^{-1/2} + p^{1-\delta/2} n^{-1/2} + \|\widehat{\Sigma}_\varepsilon(k)\|) \\ &= O_P(p^{1-\delta/2} n^{-1/2} + \|\widehat{\Sigma}_\varepsilon(k)\|). \end{aligned}$$

Now  $\|\widehat{\Sigma}_\varepsilon(k)\| \leq \|\widehat{\Sigma}_\varepsilon(k)\|_F = O_P(p n^{-1/2})$ , where  $\|\mathbf{M}\|_F = \text{trace}(\mathbf{M}\mathbf{M}')$  denotes the Frobenius norm of  $\mathbf{M}$ . Hence from (A.2),

$$(A.3) \quad \begin{aligned} \|\widehat{\Sigma}_y(k) - \Sigma_y(k)\| &= O_P(p n^{-1/2}) \quad \text{and} \\ \|\widehat{\mathbf{M}} - \mathbf{M}\| &= O_P(p^{1-\delta} \cdot p n^{-1/2}) = O_P(p^{2-\delta} n^{-1/2}). \end{aligned}$$

For the main proof, consider for  $j = 1, \dots, r$ , the decomposition

$$(A.4) \quad \begin{aligned} \hat{\lambda}_j - \lambda_j &= \hat{\gamma}_j' \widehat{\mathbf{M}} \hat{\gamma}_j - \mathbf{a}_j' \mathbf{M} \mathbf{a}_j = I_1 + I_2 + I_3 + I_4 + I_5 \\ \text{where } I_1 &= (\hat{\gamma}_j - \mathbf{a}_j)' (\widehat{\mathbf{M}} - \mathbf{M}) \hat{\gamma}_j, \quad I_2 = (\hat{\gamma}_j - \mathbf{a}_j)' \mathbf{M} (\hat{\gamma}_j - \mathbf{a}_j), \\ I_3 &= (\hat{\gamma}_j - \mathbf{a}_j)' \mathbf{M} \mathbf{a}_j, \quad I_4 = \mathbf{a}_j' (\widehat{\mathbf{M}} - \mathbf{M}) \hat{\gamma}_j, \\ I_5 &= \mathbf{a}_j' \mathbf{M} (\hat{\gamma}_j - \mathbf{a}_j). \end{aligned}$$

For  $j = 1, \dots, r$ , since  $\|\hat{\gamma}_j - \mathbf{a}_j\| \leq \|\widehat{\mathbf{A}} - \mathbf{A}\| = O_P(h_n)$  where  $h_n = p^\delta n^{-1/2}$ , and  $\|\mathbf{M}\| \leq \sum_{k=1}^{k_0} \|\Sigma_y(k)\|^2 = O_P(p^{2-2\delta})$  by (A.2), together with (A.3) we have that

$$\|I_1\|, \|I_2\| = O_P(p^{2-2\delta} h_n^2), \quad \|I_3\|, \|I_4\|, \|I_5\| = O_P(p^{2-2\delta} h_n),$$

so that  $|\hat{\lambda}_j - \lambda_j| = O_P(p^{2-2\delta} h_n) = O_P(p^{2-\delta} n^{-1/2})$ , which proves Theorem 1(i).

Now consider  $j = r + 1, \dots, p$ . Define

$$\widetilde{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\Sigma}_y(k) \Sigma_y(k)', \quad \widehat{\mathbf{B}} = (\widehat{\gamma}_{r+1}, \dots, \widehat{\gamma}_p), \quad \mathbf{B} = (\mathbf{a}_{r+1}, \dots, \mathbf{a}_p).$$

Following the same proof of Theorem 1 of [13], we can actually show that  $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(h_n)$ , so that  $\|\widehat{\gamma}_j - \mathbf{a}_j\| \leq \|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(h_n)$ .

Noting  $\lambda_j = 0$  for  $j = r + 1, \dots, p$ , consider the decomposition

$$\begin{aligned} \widehat{\lambda}_j &= \widehat{\gamma}_j' \widehat{\mathbf{M}} \widehat{\gamma}_j = K_1 + K_2 + K_3 \\ \text{where } K_1 &= \widehat{\gamma}_j' (\widehat{\mathbf{M}} - \widetilde{\mathbf{M}} - \widetilde{\mathbf{M}}' + \mathbf{M}) \widehat{\gamma}_j, \\ K_2 &= 2\widehat{\gamma}_j' (\widetilde{\mathbf{M}} - \mathbf{M}) (\widehat{\gamma}_j - \mathbf{a}_j), \\ K_3 &= (\widehat{\gamma}_j - \mathbf{a}_j)' \mathbf{M} (\widehat{\gamma}_j - \mathbf{a}_j). \end{aligned} \tag{A.5}$$

Using (A.3),

$$K_1 = \sum_{k=1}^{k_0} \|(\widehat{\Sigma}_y(k) - \Sigma_y(k))' \widehat{\gamma}_j\|^2 \leq \sum_{k=1}^{k_0} \|\widehat{\Sigma}_y(k) - \Sigma_y(k)\|^2 = O_P(p^2 n^{-1}).$$

Similarly, using (A.2) and (A.3), and  $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(h_n)$ , we can show that

$$\begin{aligned} |K_2| &= O_P(\|\widetilde{\mathbf{M}} - \mathbf{M}\| \cdot \|\widehat{\gamma}_j - \mathbf{a}_j\|) = O_P(\|\widetilde{\mathbf{M}} - \mathbf{M}\| \cdot \|\widehat{\mathbf{B}} - \mathbf{B}\|) = O_P(p^2 n^{-1}), \\ |K_3| &= O_P(\|\widehat{\mathbf{B}} - \mathbf{B}\|^2 \cdot \|\mathbf{M}\|) = O_P(p^{2-2\delta} h_n^2) = O_P(p^2 n^{-1}). \end{aligned}$$

Hence  $\widehat{\lambda}_j = O_P(p^2 n^{-1})$ , and the proof of the theorem is completed.  $\square$

**PROOF OF COROLLARY 1.** The proof of Theorem 1 of [13] has shown that (in the notation of this paper)

$$p^{2-2\delta} = O(\lambda_r).$$

But we also have

$$\lambda_r \leq \lambda_1 = \|\mathbf{M}\| \leq \sum_{k=1}^{k_0} \|\Sigma_y(k)\|^2 = O(p^{2-2\delta}),$$

where the last equality sign follows from  $\|\Sigma_y(k)\| = O(p^{1-\delta})$  in (A.2). Hence we have  $\lambda_i \asymp p^{2-2\delta}$  for  $i = 1, \dots, r$ .

Letting  $e_i = |\widehat{\lambda}_i - \lambda_i|$  for  $i = 1, \dots, r$ , we then have  $e_i = O_P(p^{2-\delta} n^{-1/2})$  from Theorem 1(i). But since  $h_n = p^\delta n^{-1/2} = o(1)$  implying that  $p^{2-\delta} n^{-1/2} = p^{2-2\delta} h_n = o(p^{2-2\delta})$ , we have  $e_i = o_P(\lambda_i)$ . Hence we must have  $\widehat{\lambda}_i \asymp \lambda_i \asymp$

$p^{2-2\delta}$  for  $i = 1, \dots, r$ . This implies that  $\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \asymp 1$  for  $j = 1, \dots, r-1$ , and together with Theorem 1(ii),

$$\widehat{\lambda}_{r+1}/\widehat{\lambda}_r = O_P(p^2 n^{-1}/p^{2-2\delta}) = O_P(p^{2\delta} n^{-1}) = O_P(h_n^2).$$

This completes the proof of the corollary.  $\square$

In the following, we use  $\sigma_j(\mathbf{M})$  to denote the  $j$ th largest singular value of a matrix  $\mathbf{M}$ , so that  $\sigma_1(\mathbf{M}) = \|\mathbf{M}\|$ . We use  $\lambda_j(\mathbf{M})$  to denote the  $j$ th largest eigenvalue of  $\mathbf{M}$ .

**PROOF OF THEOREM 2.** The first part of the theorem is actually Theorem 2 of [13]. We prove the other parts of the theorem. From equation (22) of [13], the sample lag- $k$  autocovariance matrix for  $\boldsymbol{\varepsilon}_t$  satisfies

$$(A.6) \quad \|\widehat{\boldsymbol{\Sigma}}_\epsilon(k)\| = O_P(pn^{-1}).$$

Note that (A.2) together with (A.6) implies

$$\begin{aligned} \|\widehat{\mathbf{M}} - \mathbf{M}\| &= O_P(p^{1-\delta}(p^{1-\delta/2}n^{-1/2} + pn^{-1})) \\ &= O_P(p^{2-2\delta}(p^{\delta/2}n^{-1/2} + p^\delta n^{-1})) = O_P(p^{2-2\delta}\ell_n), \end{aligned}$$

since  $\ell_n = p^{\delta/2}n^{-1/2} = o(1)$ . We also have  $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_P(\ell_n)$ , similar to the proof of Theorem 1.

With these, for  $j = 1, \dots, r$ , using decomposition (A.4), we have

$$|\widehat{\lambda}_j - \lambda_j| = O_P(\|\widehat{\mathbf{M}} - \mathbf{M}\|) = O_P(p^{2-2\delta}\ell_n) = O_P(p^{2-3\delta/2}n^{-1/2}),$$

which is Theorem 2(i). For  $j = r+1, \dots, (k_0+1)r$ , using decomposition (A.5), we have

$$\begin{aligned} K_1 &= O_P((p^{1-\delta/2}n^{-1/2} + pn^{-1})^2) = O_P(p^{2-\delta}n^{-1} + p^2n^{-2}) = O_P(p^{2-\delta}n^{-1}), \\ |K_2| &= O_P(\|\widehat{\mathbf{M}} - \mathbf{M}\| \cdot \|\widehat{\mathbf{B}} - \mathbf{B}\|) = O_P(p^{2-2\delta}\ell_n^2) = O_P(p^{2-\delta}n^{-1}), \\ |K_3| &= O_P(\|\widehat{\mathbf{B}} - \mathbf{B}\|^2 \cdot \|\mathbf{M}\|) = O_P(p^{2-2\delta}\ell_n^2) = O_P(p^{2-\delta}n^{-1}). \end{aligned}$$

Hence  $\widehat{\lambda}_j = O_P(p^{2-2\delta}\ell_n^2) = O_P(p^{2-\delta}n^{-1})$ , which is Theorem 2(ii).

For part (iii), we define

$$\mathbf{W}_y(k_0) = (\boldsymbol{\Sigma}_y(1), \dots, \boldsymbol{\Sigma}_y(k_0)), \quad \widehat{\mathbf{W}}_y(k_0) = (\widehat{\boldsymbol{\Sigma}}_y(1), \dots, \widehat{\boldsymbol{\Sigma}}_y(k_0)),$$

so that  $\mathbf{M} = \mathbf{W}_y(k_0)\mathbf{W}_y(k_0)'$  and  $\widehat{\mathbf{M}} = \widehat{\mathbf{W}}_y(k_0)\widehat{\mathbf{W}}_y(k_0)'$ . We define similarly  $\widehat{\mathbf{W}}_x(k_0)$ ,  $\widehat{\mathbf{W}}_{x\epsilon}(k_0)$ ,  $\widehat{\mathbf{W}}_{\epsilon x}(k_0)$  and  $\widehat{\mathbf{W}}_\epsilon(k_0)$ . Then we can write

$$\widehat{\mathbf{W}}_y(k_0) = M_1 + M_2 + \widehat{\mathbf{W}}_\epsilon(k_0),$$

where  $M_1 = \mathbf{A}(\widehat{\mathbf{W}}_x(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{A}') + \widehat{\mathbf{W}}_{x\epsilon}(k_0))$ ,  $M_2 = \widehat{\mathbf{W}}_{\epsilon x}(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{A}')$ . It is easy to see that

$$\text{rank}(M_1) \leq r, \quad \text{rank}(M_2) \leq k_0 r,$$

so that  $\text{rank}(M_1 + M_2) \leq (k_0 + 1)r$ . This implies that

$$\sigma_j(M_1 + M_2) = 0 \quad \text{for } j = (k_0 + 1)r + 1, \dots, p.$$

Then by Theorem 3.3.16(a) of [11], for  $j = (k_0 + 1)r + 1, \dots, p$ ,

$$\begin{aligned} \widehat{\lambda}_j &= \lambda_j(\widehat{\mathbf{M}}) = \sigma_j^2(\widehat{\mathbf{W}}_y(k_0)) \leq (\sigma_j(M_1 + M_2) + \sigma_1(\widehat{\mathbf{W}}_\epsilon(k_0)))^2 \\ &= \sigma_1^2(\widehat{\mathbf{W}}_\epsilon(k_0)) \leq \sum_{k=1}^{k_0} \|\widehat{\boldsymbol{\Sigma}}_\epsilon(k)\|^2 = O_P(p^2 n^{-2}), \end{aligned}$$

where the last equality sign follows from (A.6). This proves Theorem 2(iii).

We prove Theorem 2(ii)' now. Using Lemma 3 of [13], with the same technique as in the proof of Theorem 1 in their paper, we can write

$$(A.7) \quad \widehat{\mathbf{B}} = (\mathbf{B} + \mathbf{A}\mathbf{P})(\mathbf{I} + \mathbf{P}'\mathbf{P})^{-1/2} \quad \text{with } \|\mathbf{P}\| = O_P(\ell_n).$$

With the definition of  $\widehat{\mathbf{B}}$  as in the proof of Theorem 1, we can write  $\widehat{\lambda}_{r+1}$ , the  $(r+1)$ th largest eigenvalue of  $\widehat{\mathbf{M}}$ , as the  $(1,1)$  element of the diagonal matrix  $\widehat{\mathbf{D}} = \widehat{\mathbf{B}}'\widehat{\mathbf{M}}\widehat{\mathbf{B}}$ , where  $\widehat{\mathbf{M}}\widehat{\mathbf{B}} = \widehat{\mathbf{B}}\widehat{\mathbf{D}}$ . But from (A.7), we also have  $\mathbf{B}'\widehat{\mathbf{B}} = \mathbf{B}'(\mathbf{B} + \mathbf{A}\mathbf{P})(\mathbf{I} + \mathbf{P}'\mathbf{P})^{-1/2} = (\mathbf{I} + \mathbf{P}'\mathbf{P})^{-1/2}$ , hence

$$(\mathbf{I} + \mathbf{P}'\mathbf{P})^{1/2}\mathbf{B}'\widehat{\mathbf{M}}\widehat{\mathbf{B}} = (\mathbf{I} + \mathbf{P}'\mathbf{P})^{1/2}\mathbf{B}'\widehat{\mathbf{B}}\widehat{\mathbf{D}} = (\mathbf{I} + \mathbf{P}'\mathbf{P})^{1/2}(\mathbf{I} + \mathbf{P}'\mathbf{P})^{-1/2}\widehat{\mathbf{D}} = \widehat{\mathbf{D}}.$$

Further, by using Neumann series expansions of  $(\mathbf{I} + \mathbf{P}'\mathbf{P})^{1/2}$  and  $(\mathbf{I} + \mathbf{P}'\mathbf{P})^{-1/2}$ , we see that the largest order term of  $(\mathbf{I} + \mathbf{P}'\mathbf{P})^{1/2}\mathbf{B}'\widehat{\mathbf{M}}\widehat{\mathbf{B}}$  is contributed from  $\mathbf{B}'\widehat{\mathbf{M}}(\mathbf{B} + \mathbf{A}\mathbf{P})$ , since from (A.7) we have  $\|\mathbf{P}\| = O_P(\ell_n) = o_P(1)$ . Hence the rate of  $\widehat{\lambda}_{r+1}$  can be analyzed using the  $(1,1)$  element of  $\mathbf{B}'\widehat{\mathbf{M}}(\mathbf{B} + \mathbf{A}\mathbf{P})$ .

Some notation first. Define  $\mathbf{1}_k$  the column vector of  $k$  ones, and

$$\mathbf{E}_{r,s} = (\boldsymbol{\varepsilon}_r, \dots, \boldsymbol{\varepsilon}_s), \quad \mathbf{X}_{r,s} = (\mathbf{x}_r, \dots, \mathbf{x}_s) \quad \text{for } r \leq s.$$

Since  $k$  is finite and  $\{\boldsymbol{\varepsilon}_t\}$  and  $\{\mathbf{x}_t\}$  are stationary, for convenience in this proof we take the sample lag- $k$  autocovariance matrix for  $\{\boldsymbol{\varepsilon}_t\}$ ,  $\{\mathbf{x}_t\}$  and the cross lag- $k$  autocovariance matrix between  $\{\boldsymbol{\varepsilon}_t\}$  and  $\{\mathbf{x}_t\}$  to be respectively, for  $k > 0$ ,

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_\epsilon(k) &= n^{-1}(\mathbf{E}_{k+1,n} - (n-k)^{-1}\mathbf{E}_{k+1,n}\mathbf{1}_{n-k}\mathbf{1}_{n-k}') \\ &\quad \times (\mathbf{E}_{1,n-k} - (n-k)^{-1}\mathbf{E}_{1,n-k}\mathbf{1}_{n-k}\mathbf{1}_{n-k}')' \\ &= n^{-1}\mathbf{E}_{k+1,n}\mathbf{T}_{n-k}\mathbf{E}_{1,n-k}', \\ \widehat{\boldsymbol{\Sigma}}_x(k) &= n^{-1}\mathbf{X}_{k+1,n}\mathbf{T}_{n-k}\mathbf{X}_{1,n-k}' \end{aligned}$$

and

$$\widehat{\Sigma}_{x\epsilon}(k) = n^{-1} \mathbf{X}_{k+1,n} \mathbf{T}_{n-k} \mathbf{E}'_{1,n-k},$$

where  $\mathbf{T}_j = \mathbf{I}_j - j^{-1} \mathbf{1}_j \mathbf{1}'_j$ . Then

$$\mathbf{B}' \widehat{\mathbf{M}} (\mathbf{B} + \mathbf{A} \mathbf{P}) = \sum_{k=1}^{k_0} \mathbf{B}' \widehat{\Sigma}_y(k) \widehat{\Sigma}_y(k)' (\mathbf{B} + \mathbf{A} \mathbf{P}) = \sum_{k=1}^{k_0} \mathbf{F}_k (\mathbf{F}'_k + \mathbf{G}_k),$$

where

$$\begin{aligned} \mathbf{F}_k &= n^{-1} \mathbf{B}' \mathbf{E}_{k+1,n} \mathbf{T}_{n-k} \mathbf{X}'_{1,n-k} \mathbf{A}' + n^{-1} \mathbf{B}' \mathbf{E}_{k+1,n} \mathbf{T}_{n-k} \mathbf{E}'_{1,n-k}, \\ \mathbf{G}_k &= n^{-1} \mathbf{A} \mathbf{X}_{1,n-k} \mathbf{T}_{n-k} \mathbf{X}'_{k+1,n} \mathbf{P}' + n^{-1} \mathbf{E}_{1,n-k} \mathbf{T}_{n-k} \mathbf{X}'_{k+1,n} \mathbf{P}' \\ &\quad + n^{-1} \mathbf{A} \mathbf{X}_{1,n-k} \mathbf{T}_{n-k} \mathbf{E}'_{k+1,n} \mathbf{A} \mathbf{P}' + n^{-1} \mathbf{E}_{1,n-k} \mathbf{T}_{n-k} \mathbf{E}'_{k+1,n} \mathbf{A} \mathbf{P}'. \end{aligned}$$

Some tedious algebra (omitted here) shows that the dominating term of the above product is  $\sum_{k=1}^{k_0} n^{-2} \mathbf{B}' \mathbf{E}_{k+1,n} \mathbf{T}_{n-k} \mathbf{X}'_{1,n-k} \mathbf{X}_{1,n-k} \mathbf{T}_{n-k} \mathbf{X}'_{k+1,n} \mathbf{P}'$ . Defining  $\mathbf{c}'_{1,k} = (\mathbf{a}'_{r+1} \boldsymbol{\epsilon}_{k+1}, \dots, \mathbf{a}'_{r+1} \boldsymbol{\epsilon}_n)$  and  $\mathbf{p}_1$  the first column of  $\mathbf{P}'$ , the  $(1, 1)$  element of the said term is then

$$\begin{aligned} &\sum_{k=1}^{k_0} n^{-2} \mathbf{c}'_{1,k} \mathbf{T}_{n-k} \mathbf{X}'_{1,n-k} \mathbf{X}_{1,n-k} \mathbf{T}_{n-k} \mathbf{X}'_{k+1,n} \mathbf{p}_1 \\ &\leq \sum_{k=1}^{k_0} n^{-2} \|\mathbf{c}'_{1,k}\| \|\mathbf{p}_1\| \|\mathbf{T}_{n-k}\|^2 \|\mathbf{X}_{1,n-k}\|^2 \|\mathbf{X}_{k+1,n}\| \\ &\leq 4 \sum_{k=1}^{k_0} \|n^{-1/2} \mathbf{c}_{1,k}\| \|\mathbf{P}\| \|n^{-1/2} \mathbf{X}_{1,n-k}\|^2 \|n^{-1/2} \mathbf{X}_{k+1,n}\| \\ &= O_P(\|n^{-1/2} \mathbf{c}_{1,1}\| \cdot \ell_n \cdot p^{(3-3\delta)/2}). \end{aligned}$$

In the last line we used  $\|n^{-1/2} \mathbf{X}_{1,n-k}\| = O_P(p^{(1-\delta)/2})$ , by noting that

$$\begin{aligned} \|n^{-1/2} \mathbf{X}_{1,n-k}\|^2 &= \|n^{-1} \mathbf{X}_{1,n-k} \mathbf{X}'_{1,n-k}\| \asymp \|n^{-1} \mathbf{X}_{1,n-k} \mathbf{T}_{n-k} \mathbf{X}'_{1,n-k}\| \\ &= \|\widehat{\Sigma}_x(0)\| \leq \|\widehat{\Sigma}_x(0) - \Sigma_x(0)\| + \|\Sigma_x(0)\| \\ &= O_P(p^{1-\delta} n^{-1/2}) + O_P(p^{1-\delta}) = O(p^{1-\delta}), \end{aligned}$$

where  $\|\widehat{\Sigma}_x(0) - \Sigma_x(0)\| = O_P(p^{1-\delta} n^{-1/2})$  is from (A.1) and  $\|\Sigma_x(0)\| = O(p^{1-\delta})$  is assumed in condition (C5). With condition (C9), we can show that  $\|n^{-1/2} \mathbf{c}_{1,1}\| = O_P(1)$ , since

$$\begin{aligned} P(\|n^{-1/2} \mathbf{c}_{1,1}\| > x) &= P\left(n^{-1} \sum_{j=k+1}^n \mathbf{a}'_{r+1} \boldsymbol{\epsilon}_j \boldsymbol{\epsilon}'_j \mathbf{a}_{r+1} > x^2\right) \\ &\leq (n-k) \mathbf{a}'_{r+1} \Sigma_{\epsilon} \mathbf{a}_{r+1} / (nx^2) \leq \sigma_{\max}^2 / x^2, \end{aligned}$$



where we used the Markov inequality with  $\sigma_{\max}^2$  the maximum eigenvalue of  $\Sigma_\varepsilon$ , and the fact that  $\mathbf{a}'_{r+1}\mathbf{a}_{r+1} = 1$ . Hence the  $(1, 1)$  element of  $\mathbf{B}'\widehat{\mathbf{M}}(\mathbf{B} + \mathbf{A}\mathbf{P})$  has rate  $O_P(p^{(3-3\delta)/2}\ell_n) = O_P(p^{3/2-\delta}n^{-1/2})$ , which is also the rate of  $\widehat{\lambda}_j$  for  $j \geq r+1$ . This completes the proof of the theorem.  $\square$

We outline the proofs of Theorems 3 and 4 below. Detailed proofs can be found in the supplemental article (Lam and Yao [12]).

OUTLINE PROOF OF THEOREM 3. First, under model (4.1) and  $\mathbf{M}$  defined in (2.6), with conditions (C1)–(C4), (C5)', (C6)', we can show that the rates of the eigenvalues of  $\mathbf{M}$  are given by

$$(A.8) \quad \lambda_j \asymp \begin{cases} p^2, & \text{for } j = 1, \dots, r_1; \\ p^{2-2\delta_2}, & \text{if } \|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2}) \text{ } (c = 1) \\ p^{2-2c\delta_2}, & \text{for } j = r_1 + 1, \dots, r; \\ & \text{if } \|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}, 1/2 \leq c < 1, \text{ and} \\ & \|\mathbf{W}_1\mathbf{W}'_{21}\| \leq q\|\mathbf{W}_1\|_{\min}\|\mathbf{W}_{21}\|, 0 \leq q < 1, \\ & \text{for } j = r_1 + 1, \dots, r. \end{cases}$$

For model (4.1), and  $\mathbf{M}^*$  defined in Section 4 by  $\mathbf{y}_t^*$  in (4.2), we have

$$(A.9) \quad \lambda_j^* \asymp p^{2-2\delta_2} \quad \text{for } j = 1, \dots, r_2.$$

We cannot use Lemma 3 of [13] to prove this theorem for the one-step estimation, since the condition  $\|\mathbf{E}\| \leq \text{sep}(\mathbf{D}_1, \mathbf{D}_2)/5$  gives a restrictive condition on the growth rate of  $p$ , and also restricts the range of  $\delta_2$  allowed. Instead, we use Theorem 4.1 of [24].

Write  $\mathbf{M} = \mathbf{X}_{ij}\mathbf{D}_{ij}\mathbf{X}'_{ij}$  for  $i \neq j = 1, 2$ , where  $\mathbf{X}_{ij} = (\mathbf{A}_i\mathbf{A}_j\mathbf{B})$ ,  $\mathbf{B}$  is the orthogonal complement of  $\mathbf{A} = (\mathbf{A}_1\mathbf{A}_2)$ , and  $\mathbf{D}_{ij}$  is diagonal with  $\mathbf{D}_{ij} = \text{diag}(\mathbf{D}_i, \mathbf{D}_j, \mathbf{0})$  where  $\mathbf{D}_1$  contains  $\lambda_j$  for  $j = 1, \dots, r_1$  and  $\mathbf{D}_2$  contains  $\lambda_j$  for  $j = r_1 + 1, \dots, r$ . With  $\mathbf{E} = \widehat{\mathbf{M}} - \mathbf{M}$ , define

$$\mathbf{X}'\mathbf{E}\mathbf{X} = (\mathbf{E}_{ij}) \quad \text{for } 1 \leq i, j \leq 3,$$

where  $\mathbf{E}_{ij} = \mathbf{A}'_i\mathbf{E}\mathbf{A}_j$  if we denote  $\mathbf{B} = \mathbf{A}_3$ .

Define  $\text{sep}(\mathbf{M}_1, \mathbf{M}_2) = \min_{\lambda \in \lambda(\mathbf{M}_1), \mu \in \lambda(\mathbf{M}_2)} |\lambda - \mu|$ . If we can show that

$$(A.10) \quad \begin{aligned} & \|(\mathbf{E}_{ij}, \mathbf{E}_{i3})\| = o_P(\gamma_{ij}) \\ & \text{with } \gamma_{ij} = \text{sep}\left(\mathbf{D}_i + \mathbf{E}_{ii}, \begin{pmatrix} \mathbf{D}_j + \mathbf{E}_{jj} & \mathbf{E}_{j3} \\ \mathbf{E}_{3j} & \mathbf{E}_{33} \end{pmatrix}\right), \end{aligned}$$

then condition (4.2) in [24] is satisfied asymptotically, so that we can use their Theorem 4.1 to conclude that for  $i \neq j = 1, 2$ ,

$$(A.11) \quad \|\widehat{\mathbf{A}}_i - \mathbf{A}_i\| = O_P(\|(\mathbf{E}_{ij}, \mathbf{E}_{i3})\|/\gamma_{ij}).$$

Since we can show that  $\|\mathbf{E}_{12}\| = O_P(\|\mathbf{E}_{13}\|) = O_P(p^2 n^{-1/2})$ , we have  $\|(\mathbf{E}_{12}, \mathbf{E}_{13})\| = O_P(p^2 n^{-1/2})$ . We can also show that  $\gamma_{12} \asymp p^2$  using (A.8). Hence (A.10) is satisfied, and (A.11) implies that

$$\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\| = O_P(p^2 n^{-1/2}/p^2) = O_P(n^{-1/2}).$$

Also, we can show that  $\|\mathbf{E}_{23}\| = O_P(\|\mathbf{E}_{21}\|) = O_P(p^{2-\delta_2/2} n^{-1/2})$ , implying that  $\|(\mathbf{E}_{21}, \mathbf{E}_{23})\| = O_P(p^{2-\delta_2/2} n^{-1/2})$ . We can also show that  $\gamma_{21} \asymp p^{2-2c\delta_2}$  using (A.8), provided  $p^{c\delta_2} n^{-1/2} \rightarrow 0$ . Hence (A.10) is satisfied since we assumed  $\nu_n \rightarrow 0$ , and so (A.11) implies that

$$\|\widehat{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(p^{2-\delta_2/2} n^{-1/2}/p^{2-2c\delta_2}) = O_P(p^{(2c-1/2)\delta_2} n^{-1/2}) = O_P(\nu_n),$$

which completes the proof for the one-step estimation.

For the two-step estimation, write  $\mathbf{M}^* = (\mathbf{A}_2 \mathbf{B}^*) \mathbf{D}^* (\mathbf{A}_2 \mathbf{B}^*)'$ , where  $\mathbf{B}^*$  is the orthogonal complement of  $\mathbf{A}_2$ , and  $\mathbf{D}^*$  is diagonal with  $\mathbf{D}^* = \text{diag}(\mathbf{D}_2^*, \mathbf{0})$ . The matrix  $\mathbf{D}_2^*$  contains  $\lambda_j^*$  for  $j = 1, \dots, r_2$ , so that (A.9) implies  $\text{sep}(\mathbf{D}_2^*, \mathbf{0}) \asymp p^{2-2\delta_2}$ .

We can show that  $\|\mathbf{E}^*\| = \|\widehat{\mathbf{M}}^* - \mathbf{M}^*\| = O_P(p^{2-2\delta_2} \kappa_n)$ , hence  $\|\mathbf{E}^*\| = o_P(\text{sep}(\mathbf{D}_2^*, \mathbf{0}))$ , as  $\kappa_n \rightarrow 0$ . Hence we can use Lemma 3 of [13] to conclude that

$$\|\widetilde{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(\|\mathbf{E}_{21}^*\| / \text{sep}(\mathbf{D}_2^*, \mathbf{0})).$$

Since we can show that  $\|\mathbf{E}_{21}^*\| = O_P(p^{2-3\delta_2/2} n^{-1/2})$ , we then have

$$\|\widetilde{\mathbf{A}}_2 - \mathbf{A}_2\| = O_P(p^{2-3\delta_2/2} n^{-1/2}/p^{2-2\delta_2}) = O_P(p^{\delta_2/2} n^{-1/2}),$$

which completes the proof of the theorem.  $\square$

To prove Theorem 4, we need two lemmas first.

LEMMA 1. *Under the same conditions and notations of Theorem 3, the following assertions hold:*

- (i) For  $j = 1, \dots, r_1$ ,  $|\widehat{\lambda}_j - \lambda_j| = O_P(p^2 n^{-1/2})$ .
- (ii) For  $j = r_1 + 1, \dots, r$ ,  $|\widehat{\lambda}_j - \lambda_j| = O_P(p^2(n^{-1/2} + \nu_n^2))$  provided  $\nu_n \rightarrow 0$ ,  $p^{c\delta_2} n^{-1/2} \rightarrow 0$ .
- (iii) For  $j = r + 1, \dots, p$ ,  $\widehat{\lambda}_j = O_P(p^2 \nu_n^2)$ , provided  $\nu_n \rightarrow 0$ ,  $p^{c\delta_2} n^{-1/2} \rightarrow 0$ .
- (iv) For  $j = 1, \dots, r_2$ ,  $|\widehat{\lambda}_j^* - \lambda_j^*| = O_P(p^{2-2\delta_2} \kappa_n)$ .
- (v) For  $j = r_2 + 1, \dots, p$ ,  $\widehat{\lambda}_j^* = O_P(p^{2-2\delta_2} \kappa_n^2)$ .
- (vi) For  $j = (k_0 + 1)r + 1, \dots, p$ ,  $\widehat{\lambda}_j, \widehat{\lambda}_j^* = O_P(p^2 n^{-2}) = O_P(p^{2-2\delta_2} \kappa_n^4)$ .
- (iii)' If in addition condition (C9) holds, then for  $j = r + 1, \dots, p$ ,  $\widehat{\lambda}_j = O_P(p^{3/2} \nu_n)$ , provided  $\nu_n \rightarrow 0$ ,  $p^{c\delta_2} n^{-1/2} \rightarrow 0$ .

The proof of this lemma is left in the supplementary materials for this paper. Together with (A.8) and (A.9), we have the following lemma.

LEMMA 2. *Let conditions (C1)–(C4), (C5)', (C6)', (C7) and (C8) hold. Then as  $n, p \rightarrow \infty$  with  $n = O(p)$ , and with  $\nu_n, \kappa_n \rightarrow 0$  the same as in Theorem 3 and  $p^{c\delta_2}n^{-1/2} \rightarrow 0$ , we have*

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \begin{cases} \asymp 1, & j = 1, \dots, r_1 - 1; \\ = O_P(n^{-1/2} + \nu_n^2 + p^{-2\delta_2}), & j = r_1, \text{ if } \|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2}) \\ & (c = 1); \\ = O_P(n^{-1/2} + \nu_n^2 + p^{-2c\delta_2}), & j = r_1, \text{ if } \|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2} \\ & \text{for } 1/2 \leq c < 1, \text{ and} \\ & \|\mathbf{W}_1 \mathbf{W}'_{21}\| \leq q \|\mathbf{W}_1\|_{\min} \|\mathbf{W}_{21}\| \\ & \text{for } 0 \leq q < 1. \end{cases}$$

Furthermore, if  $\|\mathbf{W}_{21}\|_{\min} = o(p^{1-\delta_2})$  and  $p^{5\delta_2/2}n^{-1/2} \rightarrow 0$ , we have

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \begin{cases} \asymp 1, & j = r_1 + 1, \dots, r - 1; \\ = O_P(p^{2\delta_2}\nu_n^2), & j = r; \\ = O_P(p^{2\delta_2-1/2}\nu_n), & j = r, \text{ and condition (C9) holds.} \end{cases}$$

If  $\|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}$  for  $1/2 \leq c < 1$ ,  $\|\mathbf{W}_1 \mathbf{W}'_{21}\| \leq q \|\mathbf{W}_1\|_{\min} \|\mathbf{W}_{21}\|$  for  $0 \leq q < 1$ , and  $p^{(3c-1/2)\delta_2}n^{-1/2} \rightarrow 0$ , we have

$$\widehat{\lambda}_{j+1}/\widehat{\lambda}_j \begin{cases} \asymp 1, & j = r_1 + 1, \dots, r - 1; \\ = O_P(p^{2c\delta_2}\nu_n^2), & j = r; \\ = O_P(p^{2c\delta_2-1/2}\nu_n), & j = r, \text{ and condition (C9) holds.} \end{cases}$$

For the two-step procedure, let conditions (C1)–(C4), (C5)', (C6)', (C7) and (C8) hold and  $n = O(p)$ . Then we have

$$\widehat{\lambda}_{j+1}^*/\widehat{\lambda}_j^* \begin{cases} \asymp 1, & j = 1, \dots, r_2 - 1; \\ = O_P(\kappa_n^2), & j = r_2. \end{cases}$$

PROOF. We only need to find the asymptotic rate for each  $\widehat{\lambda}_j$  and  $\widehat{\lambda}_j^*$ . The rate of each ratio can then be obtained from the results of Lemma 1.

For  $j = 1, \dots, r_1$ , from Lemma 1,  $\|\widehat{\lambda}_j - \lambda_j\| = O_P(p^2 n^{-1/2}) = o_P(\lambda_j)$ , and hence  $\widehat{\lambda}_j \asymp \lambda_j \asymp p^2$  from (A.8).

Consider the case  $\|\mathbf{W}_{21}\|_{\min} \asymp p^{1-c\delta_2}$ . For  $j = r_1 + 1, \dots, r$ , since  $|\widehat{\lambda}_j - \lambda_j| = O_P(p^2(n^{-1/2} + \nu_n^2))$ , we have  $\widehat{\lambda}_j \leq \lambda_j + O_P(p^2(n^{-1/2} + \nu_n^2)) = O_P(p^{2-2c\delta_2} + p^2\nu_n^2 + p^2n^{-1/2})$ , and hence

$$\widehat{\lambda}_{r_1+1}/\widehat{\lambda}_{r_1} = O_P((p^{2-2c\delta_2} + p^2\nu_n^2 + p^2n^{-1/2})/p^2) = O_P(n^{-1/2} + \nu_n^2 + p^{-2c\delta_2}).$$

The other case is proved similarly.

For  $j = r_1 + 1, \dots, r$ , to make sure  $\widehat{\lambda}_j$  will not be zero or close to zero, we need

$$|\widehat{\lambda}_j - \lambda_j| = O_P(p^2(n^{-1/2} + \nu_n^2)) = o_P(\lambda_j),$$

where  $\lambda_j \asymp p^{2-2c\delta_2}$  as in (A.8). Hence we need  $p^2(n^{-1/2} + \nu_n^2) = o(p^{2-2c\delta_2})$ , which is equivalent to the condition  $p^{(3c-1/2)\delta_2}n^{-1/2} \rightarrow 0$ . With this condition satisfied, then  $\hat{\lambda}_j \asymp \lambda_j \asymp p^{2-2c\delta_2}$  for  $j = r_1 + 1, \dots, r$ . Since  $\hat{\lambda}_j = O_P(p^2\nu_n^2)$  for  $j = r + 1, \dots, p$ , we then have

$$\hat{\lambda}_{r+1}/\hat{\lambda}_r = O_P(p^2\nu_n^2/p^{2-2c\delta_2}) = O_P(p^{2c\delta_2}\nu_n^2).$$

All other rates can be proved similarly, and thus are omitted.  $\square$

**PROOF OF THEOREM 4.** With Lemma 2, Theorem 4(i) is obvious. For Theorem 4(ii), note that the range of  $\delta_2$  and the rates given in the theorem ensure that  $n^{-1/2} + \nu_n^2 + p^{-2c\delta_2} = o(p^{2c\delta_2-1/2}\nu_n) = o(p^{2c\delta_2}\nu_n^2)$ . Hence Lemma 2 implies a better rate of convergence for  $\hat{\lambda}_{r_1+1}/\hat{\lambda}_{r_1}$  no matter whether condition (C9) holds or not. We can use a similar argument to prove part (iii), and details are thus omitted.  $\square$

**Acknowledgments.** We are grateful to the Joint Editor Professor Peter Bühlmann, the Associate Editor and the two referees for their helpful comments and suggestions.

## SUPPLEMENTARY MATERIAL

**Detailed proofs of Theorems 3 and 4** (DOI: [10.1214/12-AOS970SUPP](https://doi.org/10.1214/12-AOS970SUPP); .pdf). The document contains detailed proofs of Theorem 3 and 4 in the paper.

## REFERENCES

- [1] ANDERSON, T. W. (1963). The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika* **28** 1–25. [MR0165648](#)
- [2] BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. [MR1926259](#)
- [3] BAI, J. and NG, S. (2007). Determining the number of primitive shocks in factor models. *J. Bus. Econom. Statist.* **25** 52–60. [MR2338870](#)
- [4] BATHIA, N., YAO, Q. and ZIEGELMANN, F. (2010). Identifying the finite dimensionality of curve time series. *Ann. Statist.* **38** 3352–3386. [MR2766855](#)
- [5] BRILLINGER, D. R. (1981). *Time Series: Data Analysis and Theory*, 2nd ed. Holden-Day, Oakland, CA. [MR0595684](#)
- [6] CHAMBERLAIN, G. and ROTHCHILD, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. [MR0736050](#)
- [7] FAN, J. and YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York. [MR1964455](#)
- [8] FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics* **82** 540–554.
- [9] HALLIN, M. and LIŠKA, R. (2007). Determining the number of factors in the general dynamic factor model. *J. Amer. Statist. Assoc.* **102** 603–617. [MR2325115](#)

- [10] HANNAN, E. J. (1970). *Multiple Time Series*. Wiley, New York. [MR0279952](#)
- [11] HORN, R. A. and JOHNSON, C. R. (1991). *Topics in Matrix Analysis*. Cambridge Univ. Press, Cambridge. [MR1091716](#)
- [12] LAM, C. and YAO, Q. (2012). Supplement to “Factor modeling for high-dimensional time series: Inference for the number of factors.” DOI:[10.1214/12-AOS970SUPP](#).
- [13] LAM, C., YAO, Q. and BATHIA, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika* **98** 901–918.
- [14] LÜTKEPOHL, H. (1993). *Introduction to Multiple Time Series Analysis*, 2nd ed. Springer, Berlin. [MR1239442](#)
- [15] PAN, J., PEÑA, D., POLONIK, W. and YAO, Q. (2011). Modelling multivariate volatilities via common factors. Available at <http://stats.lse.ac.uk/q.yao/qyao.links/paper/pppy.pdf>.
- [16] PAN, J. and YAO, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95** 365–379. [MR2521589](#)
- [17] PÉCHÉ, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probab. Theory Related Fields* **143** 481–516. [MR2475670](#)
- [18] PEÑA, D. and BOX, G. E. P. (1987). Identifying a simplifying structure in time series. *J. Amer. Statist. Assoc.* **82** 836–843. [MR0909990](#)
- [19] PEÑA, D. and PONCELA, P. (2006). Nonstationary dynamic factor analysis. *J. Statist. Plann. Inference* **136** 1237–1257. [MR2253761](#)
- [20] PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press, New York.
- [21] PRIESTLEY, M. B., RAO, T. S. and TONG, H. (1974). Applications of principal component analysis and factor analysis in the identification of multivariable systems. *IEEE Trans. Automat. Control* **19** 703–704.
- [22] REINSEL, G. C. (1997). *Elements of Multivariate Time Series Analysis*, 2nd ed. Springer, New York. [MR1451875](#)
- [23] SHAPIRO, D. E. and SWITZER, P. (1989). Extracting time trends from multiple monitoring sites. Technical Report 132, Dept. Statistics, Stanford Univ.
- [24] STEWART, G. W. (1973). Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Rev.* **15** 727–764. [MR0348988](#)
- [25] SWITZER, P. and GREEN, A. A. (1984). Min/Max autocorrelation factors for multivariate spatial imagery. Technical Report 6, Dept. Statistics, Stanford Univ.
- [26] TAO, M., WANG, Y., YAO, Q. and ZOU, J. (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *J. Amer. Statist. Assoc.* **106** 1025–1040.
- [27] TIAO, G. C. and TSAY, R. S. (1989). Model specification in multivariate time series (with discussion). *J. Roy. Statist. Soc. Ser. B* **51** 157–213. [MR1007452](#)
- [28] WANG, H. (2010). Factor profiling for ultra high dimensional variable selection. Available at <http://ssrn.com/abstract=1613452>.

DEPARTMENT OF STATISTICS  
 LONDON SCHOOL OF ECONOMICS  
 LONDON, WC2A 2AE  
 UNITED KINGDOM  
 E-MAIL: [c.lam2@lse.ac.uk](mailto:c.lam2@lse.ac.uk)  
[q.yao@lse.ac.uk](mailto:q.yao@lse.ac.uk)